Déjà vu – Réjà vu

On knowledge-based approaches linking ligand and target information to bioactivity

PROEFSCHRIFT

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op dinsdag 8 januari 2013

klokke 11:15 uur

door

Gerard Jacob Pieter van Westen

geboren te Leiden in 1983

Promotiecommissie

Promotores:	Prof. Dr. A.P. IJzerman
	Prof. Dr. H.W.T. van Vlijmen
Co-Promotor:	Dr. A. Bender
Overige leden:	Prof. Dr. M. Danhof
	Prof. Dr. J.N. Kok
	Prof. Dr. T. Hankemeier
	Dr. C. de Graaf

ISBN-13: 978-94-6203-261-3

The research described in this thesis was performed at the Division of Medicinal Chemistry of the Leiden/Amsterdam Center for Drug Research, Leiden University (Leiden, The Netherlands). This research was conducted as part of cooperation between the LACDR and Tibotec BVBA (Belgium) and funded by Tibotec BVBA.

This thesis was printed by Wöhrmann Print Service (Zutphen, The Netherlands)

Contents

About the title:
Déjà vu; the feeling that one has seen or experienced a certain event before.
Réjà vu; the feeling that one will see or experience something again (introduced by *T. Pratchett; Pyramids Discworld Series.1989.* ***7****. Corgi Publishers*).

Cover images:
The front cover image was obtained after text mining the full text of chapters 1 through 8 of this thesis (excluding the figure and table legends). The size of the word indicates its frequency throughout the text. The figure was made using the java based 'IBM Wordcloud' applet.

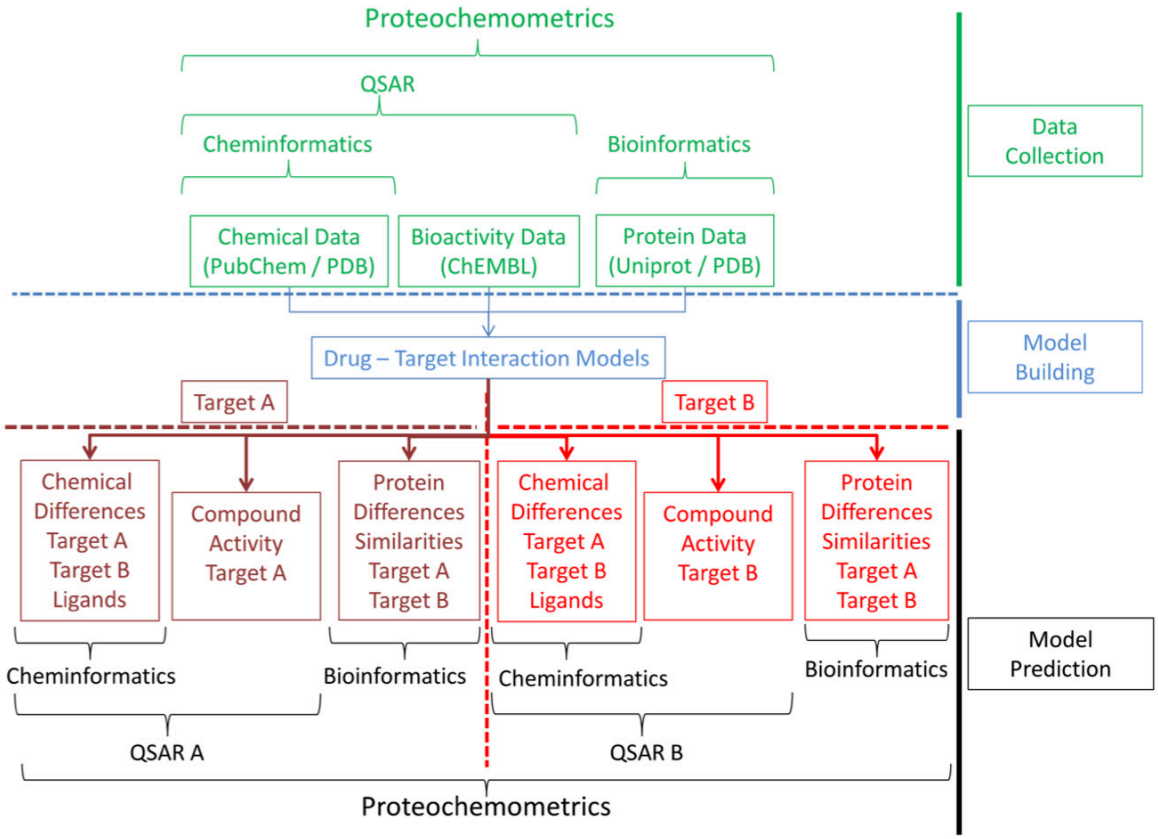
The back cover images are scatter plots of Anscombe's Quartet. These data series all produce identical values for the most common regression validation parameters yet only when analyzing the actual plots their completely different nature becomes apparent. They serve as a warning to anyone trying to create structure-activity relationships. *F.J. Anscombe; Graphs in Statistical Analysis. The American Statistician; 1973* ***27*** *(1): 17-21*.

Human beings, who are almost unique in having the ability to learn from the experience of others, are also remarkable for their apparent disinclination to do so.

Douglas Adams, "Last Chance to See", 1990, Pan Books

Chapter 1

General Introduction



Contents

1.1 About this thesis.

This thesis focuses on computational approaches able to combine data from different disciplines that are relevant in medicinal chemistry and drug discovery. These different disciplines, or data sources, are: Chemistry, Biology and Bioactivity and will be further explained below. The underlying rationale is that these disciplines are *complementary to* each other rather than *substitutes for* each other. Therefore we expect models created on data from the combination of these disciplines to be more robust than models created from data obtained from a single discipline.

1.2 Chemistry

1.2.1 Chemicals and Man. Chemicals have long since been recognized to be able to directly affect human beings. For example, at the turn of the 19th and 20th century Hans H. Meyer and Charles E. Overton both published a similar theory stating that the narcotic potency of an anesthetic can be predicted from its solubility in oil.[1, 2] While this is a very crude relationship, it can be considered one of the first attempts to correlate chemical features ("solubility") of a compound with its biological activity ("narcotic potency").

1.2.2 Small Molecules. In the 20th century a specific class of chemicals has become dominant as a source of drugs in pharmaceutical research, this class is called 'small molecules'. Small molecules are chemical compounds that are, as the name has it, relatively small. Often this size is expressed as molecular weight, where a molecular weight of approximately 500 Dalton or less is considered to be small while sometimes a limit of 800 Dalton is considered.[3] They are also required to be organic and are often relatively simple to synthesize. It is these properties that make them suitable as drugs and therefore they have been the major focus of pharmaceutical companies and medicinal chemistry research.

1.2.3 Chemical Space. With the majority of chemists focusing on the production of either novel small molecules or analogues of existing small molecules, it is not surprising that the total number of known small molecules has risen tremendously over the last years. However, the total amount of known molecules (chemical space) is not even approaching a perceptible fraction of the total number of possible small molecules (estimated at 10^{60} compounds).[4, 5] Nonetheless, as structures and properties of compounds can be stored electronically with relative ease; so called virtual libraries (collections of millions of small molecules that are theoretically possible) have appeared and can serve as a source of ideas for small molecules that are actually synthesized.

Simultaneously with virtual libraries, public databases have materialized. Public databases are similar to virtual libraries since both databases and virtual libraries contain electronically stored molecules. However, public databases differ from virtual libraries as databases contain molecules that *have been* synthesized and sometimes tested for biological activity. One of the largest free databases containing small molecules is Pubchem.[6]

1.2.4 Molecular Similarity. With the appearance of virtual libraries, a need arose to quantify their similarity (how similar is compound *A* to compound *B*). This principle is demonstrated in **Figure 1.1**. For any pair of compounds the similarity can be quantified between 0 and 1 based on the presence of absence of chemical features. When these compounds are aligned along the edges of a matrix, clusters of similar compounds appear. Given a molecule that exhibits sought properties, molecules that have similar properties can be included in the search for novel drugs. Likewise this similarity measure can be extended to three compounds or an entire virtual library.

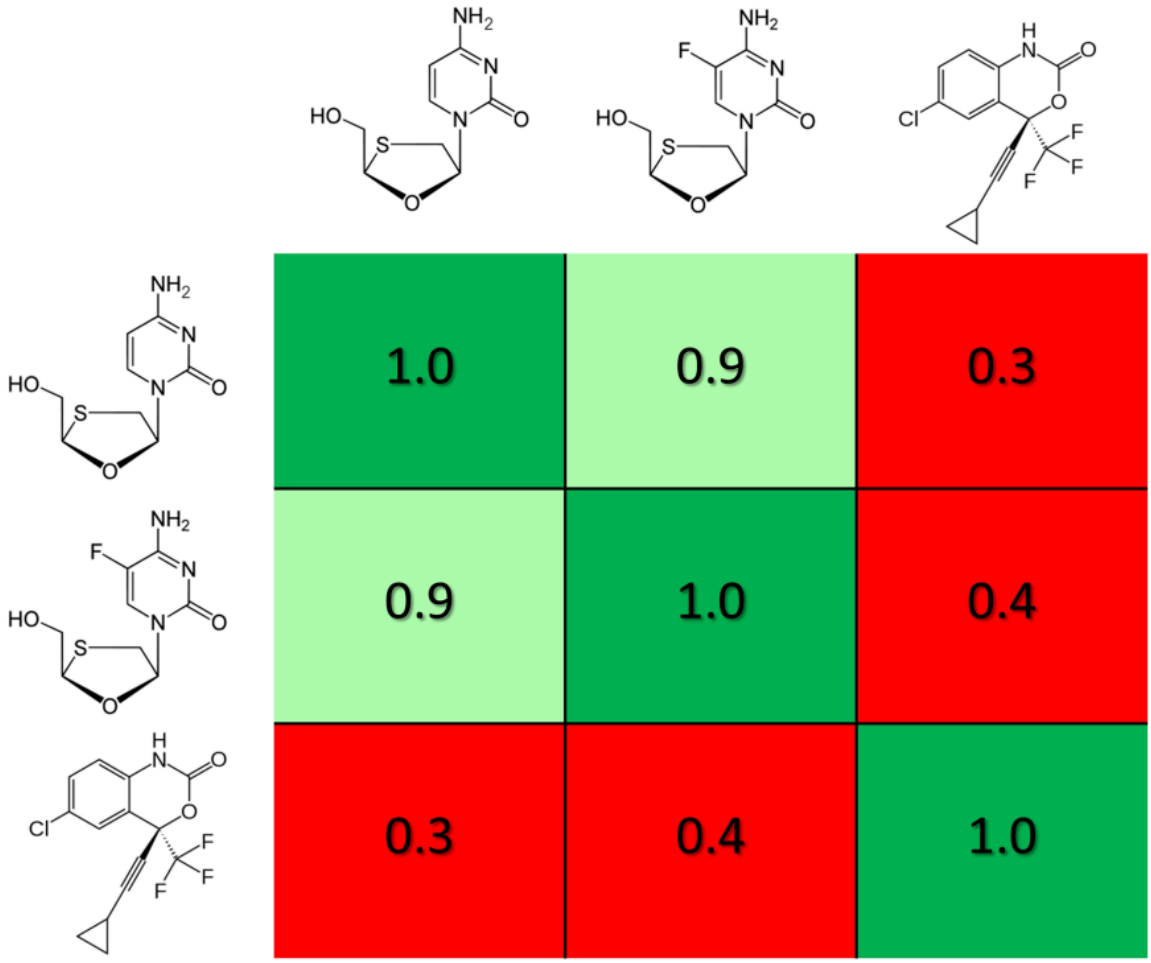


Figure 1.1: The concept of molecular similarity. For any pair of compounds the similarity can be quantified between 0 and 1 based on the presence of absence of chemical features. When these compounds are aligned along the edges of a matrix, clusters of similar compounds appear (green squares). Given a molecule that exhibits sought properties, molecules that have similar properties can be included in the search for novel drugs, while compounds that are different (red squares) are avoided.

1.3 Biology

1.3.1 Genomics. Like the field of chemistry, the field of biology, more specific the field of molecular biology, has flourished over the late 20th and early 21st centuries. Molecular biology has its roots in genetics and biochemistry. The field studies data gathered at a genetic level about gene expression and maps possible functions of these genes onto proteins. The large scale study of genetic data is also known as genomics, the study of an entire genome of an organism. The advent of genomics from molecular biology has introduced numerous techniques for large scale data storage, manipulation and data mining, the process of pattern discovery in large unsorted data sets. Computational genomics approaches use computational analysis methods to obtain these goals.

1.3.2 Proteomics. Proteomics is the large scale study of proteins consisting of experimental procedures, large scale data collection and much more. Proteomics links data gathered using genomics to proteins. In the scope of the current thesis proteomics is defined much narrower. Proteomics is of interest as computational approaches have been developed to process the large data sets that are produced on a routine basis, much like in Genomics. One of the largest free databases containing protein sequence information is Uniprot.[7] In addition to sequence information, structural information, elucidating the three dimensional structure of proteins, is also publically available. Structural information is stored in the Protein Data Bank (PDB).[8] It is both sequence information and structural information we are interested in within the scope of this thesis.

1.3.3 (Drug) Target Space. Chemistry provides a framework wherein research on small molecule drugs takes place, chemical space. Likewise, the output of genomics and proteomics provides a framework wherein research on novel proteins that can be of interest in medicinal chemistry takes place, the so-called target space. Both spaces are complementary and the advent of large scale public databases has made it possible to mine the data sets that underlie them.

1.3.4 Protein Similarity. Similar to the concept of molecular similarity (see **1.2.4**), a concept of protein similarity exists. Protein similarity can be defined on many different levels, from similarity of the three dimensional structure of two or more proteins, to the similarity between a certain region of interest which can be present on two or more proteins. This region of interest can for instance be the location where small molecules bind to the protein ('binding pocket'). The latter definition of protein similarity will be used throughout this thesis. An example is given in **Figure 1.2** using three hypothetical three amino acid peptides (but this approach can easily be up scaled to full proteins). Here, for each pair of peptides the similarity was quantified based on the physicochemical properties of the side chains and the peptides that are most similar cluster together. The rationale is as follows: given a protein that is of interest due to a specific function it performs, similar proteins can be identified which might also be capable of performing a similar function and can hence be also of interest.

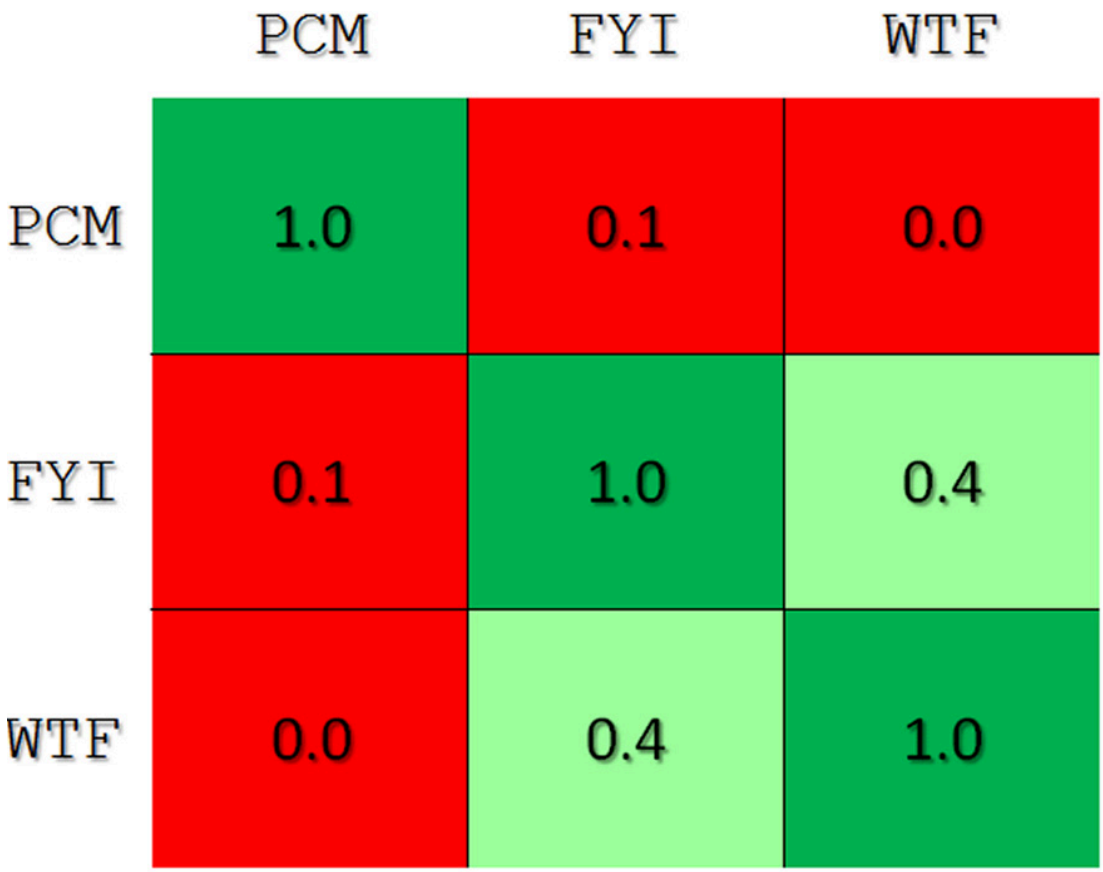


Figure 1.2: The concept of protein similarity. For any pair of peptides the similarity can for instance be quantified between 0 and 1 based on the physicochemical characteristics of the amino acid side chains. When these peptides are aligned along the edges of a matrix, clusters of similar peptides appear (green squares). Given a protein that is of interest due to a specific function it performs, similar proteins can be identified which might also be capable of performing a similar function and can hence be also of interest while dissimilar proteins (red squares) are avoided.

1.4 Bioactivity

1.4.1 Chemistry and Biology. Bioactivity is information about the effects of chemicals on living organisms. In this thesis we will focus on the effect of small molecules on proteins, bioactivity quantifies these effects. The combination of the study of chemical space with the study of target space is what makes the rational development of bioactive compounds possible. Bioactivity data has joined the ranks of molecular biology data and chemical data as publically available information. Recently bioactivity data has been included in Pubchem moreover, there is the advent of ChEMBL,[6, 9] a database completely focused on bioactivity of small molecules.

1.4.2 Exponential Data Growth. As public databases have become available, their use has also become commonplace. This has led to a nearly exponential growth in the size of the explored regions of chemical space and target space. This growth will be illustrated using the PDB over the course of its existence as a reference but is equally relevant for other databases (**Figure 1.3A**). More and more data presents scientist with the unique opportunity to start data mining for specific or global bioactivity by linking bioactivity data from sets of related targets and hence possible sets of related molecules that interact with these targets ("ligands"). Effectively, the increase in data enables rational design of desired bioactivity profiles. However, traditional methods are focused on a single target and small analogue series and are ill equipped to mine data on sets this size.

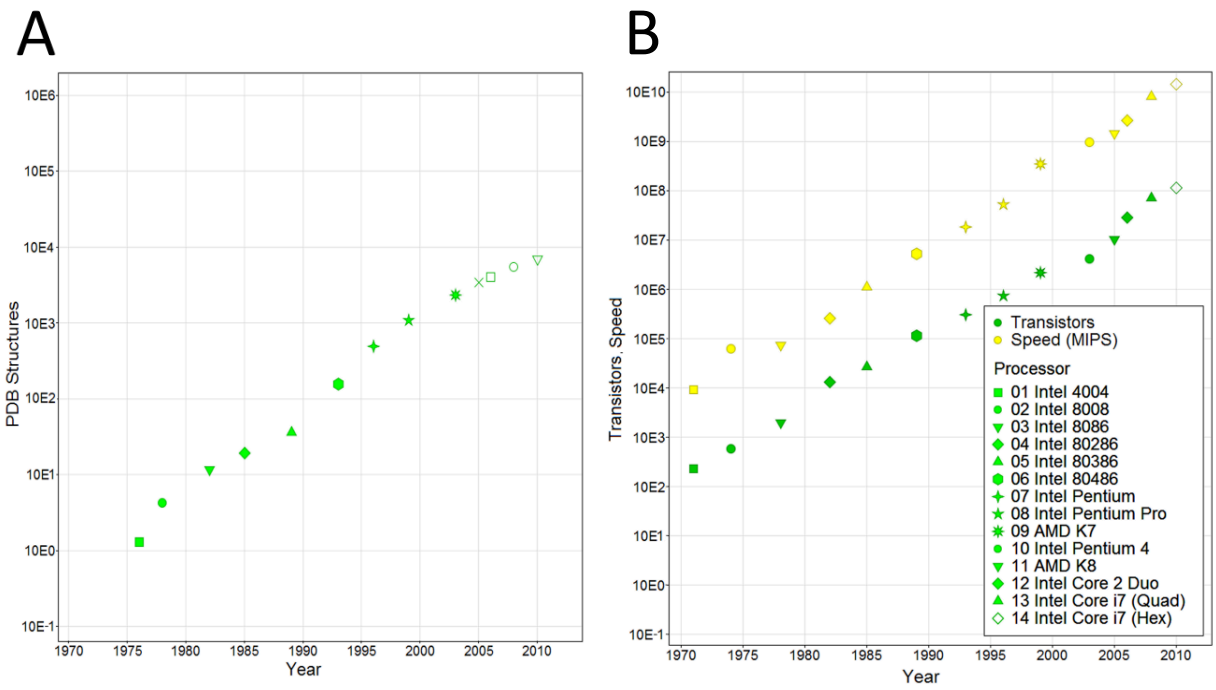


Figure 1.3: Data growth and processing power growth (A). The amount of structures available (y-axis) in the Protein Data Bank [8] from 1976 until 2010 (x-axis)(B). The increase in computing power (in Million Instructions Per Second) and transistor amount [10-14] (y-axis) of CPUs in desktop computers. Both Y-axes are drawn on a logarithmic scale.

1.5 Exponential Computational Power Growth

1.5.1 Smaller and smaller. The exponential growth in data drives a need for data analysis and fuels drug research. However, an exponential growth in the number of scientists cannot be sustained, [15, 16] neither can an exponential growth in research budget.[16, 17] Hence the increase in data analysis capacity needs to come from a different source. It is here that the exponential growth of computing power, available to standard desktop PCs, can prove instrumental (**Figure 1.3B**). One of the main driving forces behind the increase of transistors on a single CPU, hence the increase in speed is the large decrease in minimal feature width. This started as large as 10 µm in 1971 in the Intel 4004, down to 32 nm in the 2010 Intel Core i7 (Hexacore) (**Figure 1.4**).[18] Also shown is the average size of an HIV Virion particle (145 nm),[19] Staphylococcus Aureus bacterium (800 nm) [20] and a red blood cell (90 µm).[21]

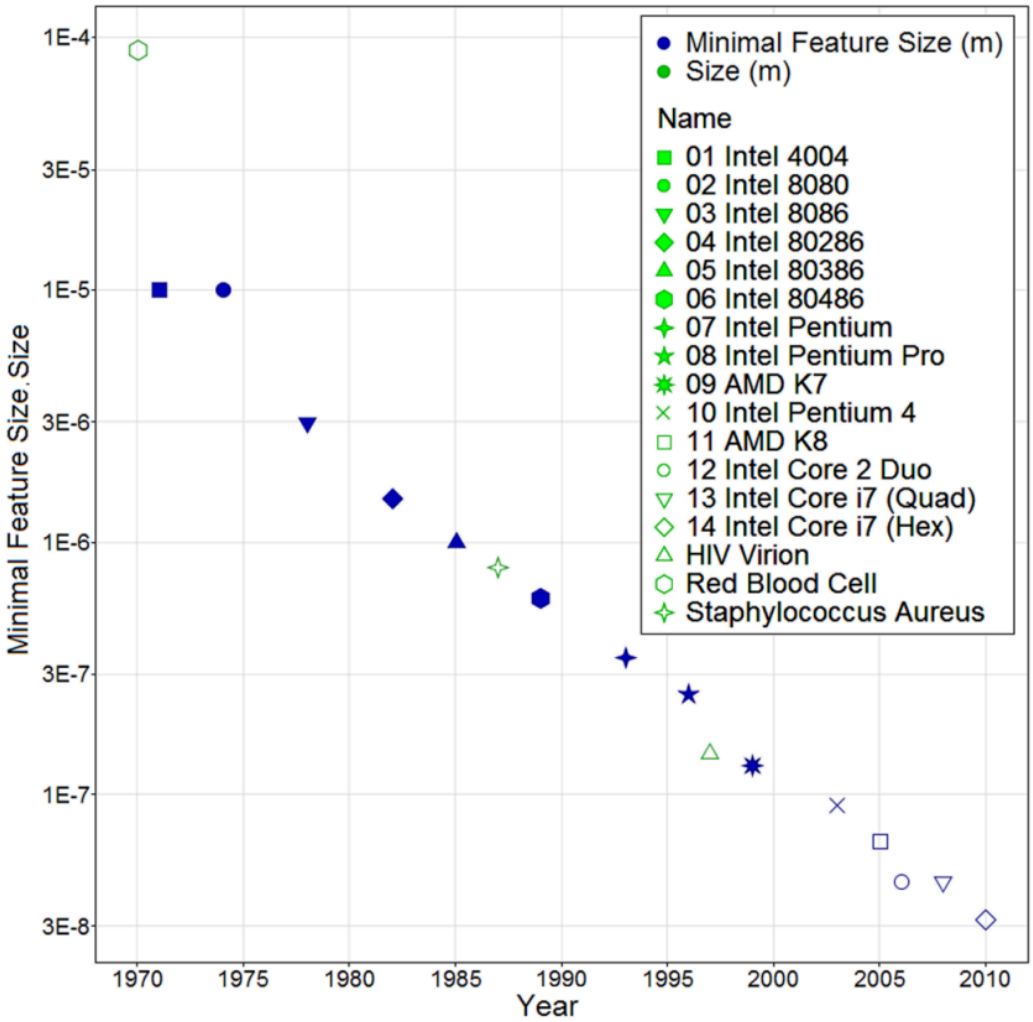


Figure 1.4: Minimal feature width on integrated circuits since 1971 until 2010.[10] Also shown are the average sizes of an HIV Virion particle, a human red blood cell, and a Staphylococcus Aureus bacterium.

1.6 Bioinformatics and Cheminformatics

1.6.1 Computers in Medicinal Chemistry. The use of computers in drug research is not novel. Their role has been invaluable in several parts of the drug design process. Among these is the creation of statistical models that explain the interaction of a small molecule to a target, so called Quantitative Structure-Activity Relationships (QSARs), but computers are also involved in X-ray crystallography. More recently the field of bioinformatics, [7, 22, 23] and cheminformatics have gained solid ground.[24] Both techniques deal with processing large amounts of data (hence the informatics suffix). These concepts will be explained below.

1.6.2 Bioinformatics. Bioinformatics combines biological information (nucleotide sequences, amino acid sequences) with computational techniques and here primary applications are storing, retrieving, and evaluating ("mining") data (**Figure 1.5**). Therefore Bioinformatics can be seen as the informatics extension to (molecular) biology and one of the major tools to navigate target space. In order for bioinformatics approaches to function, the data available needs to be transformed into information that is accessible for the computer. In practice this involves structuring the data, standardizing measurements, standardizing descriptive parameters and most important of all removing false information or noise.

After this information has been processed, the output from the computer needs to be interpreted to make it useful information accessible for scientists involved in a research project. This involves creation of plots illustrating possible correlations / inverse correlations or the use of specialized tools. Bioinformatics output can lead to insights about details of cellular modifications that are applied to proteins, or can lead to discovery of similarities between certain sets of proteins.

A simplified example application of a Bioinformatics approach to a dataset consisting of two targets (proteins) is shown in **Figure 1.5**. The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Bioinformatics can predict protein differences, protein similarities, and protein properties in general. Data sources in this case can be databases like Uniprot or the PDB.

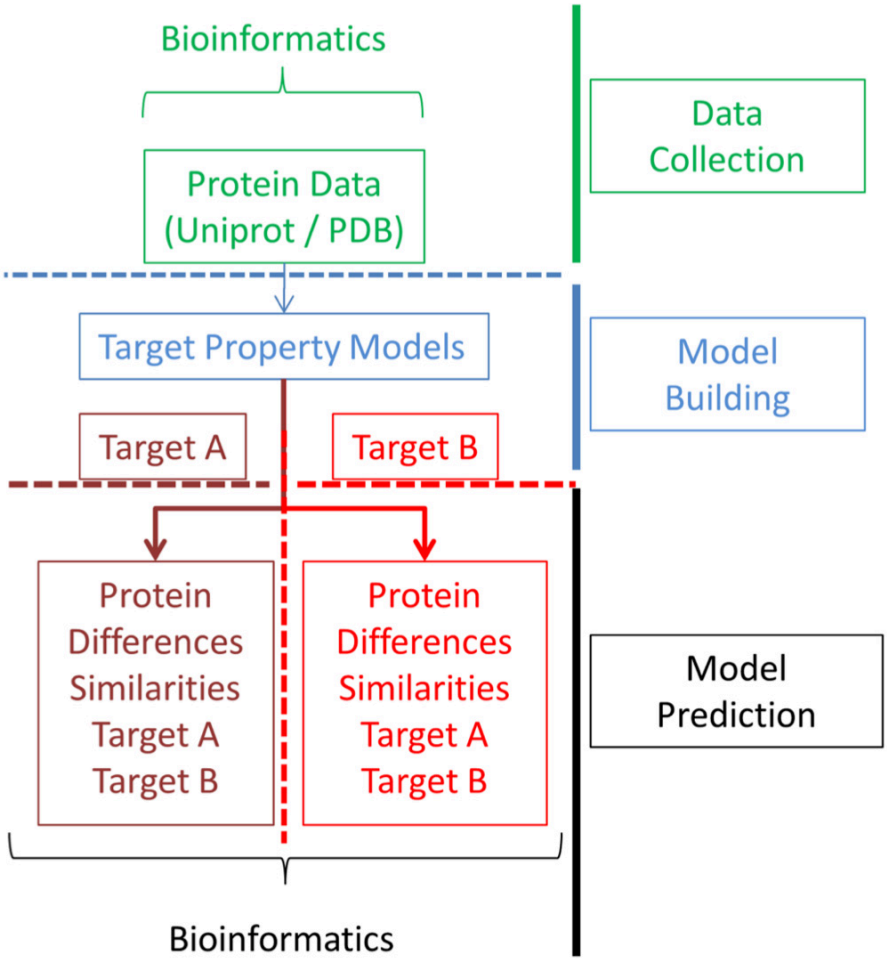


Figure 1.5: Simplified schematic overview of a bioinformatics project applied to a dataset consisting of two targets (proteins). The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

1.6.3 Cheminformatics. Cheminformatics combines chemical information with computational techniques and primary applications are storing, retrieving, and data mining of chemical information. Cheminformatics can therefore be seen as the informatics extension to chemistry and one of the major tools to navigate chemical space. Like in bioinformatics, data needs to be structured, standardized and cleared of noise.

Cheminformatics encounters hurdles very similar to bioinformatics, namely data interpretation and presentation of data in an organized fashion to other scientists involved in a research project. Hence specialized tools have been developed to retrieve compounds with desired properties or to visualize a correlation that might not be apparent on first sight.

A simplified example application of a Cheminformatics approach to a dataset consisting of two targets is shown in **Figure 1.6**. This scheme also distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Cheminformatics can predict chemical differences, chemical similarities, and chemical properties of ligands in general. Data sources in this case can be databases like Pubchem or the PDB. Also shown is the application of statistical QSAR on the same data set, please see below for further details.

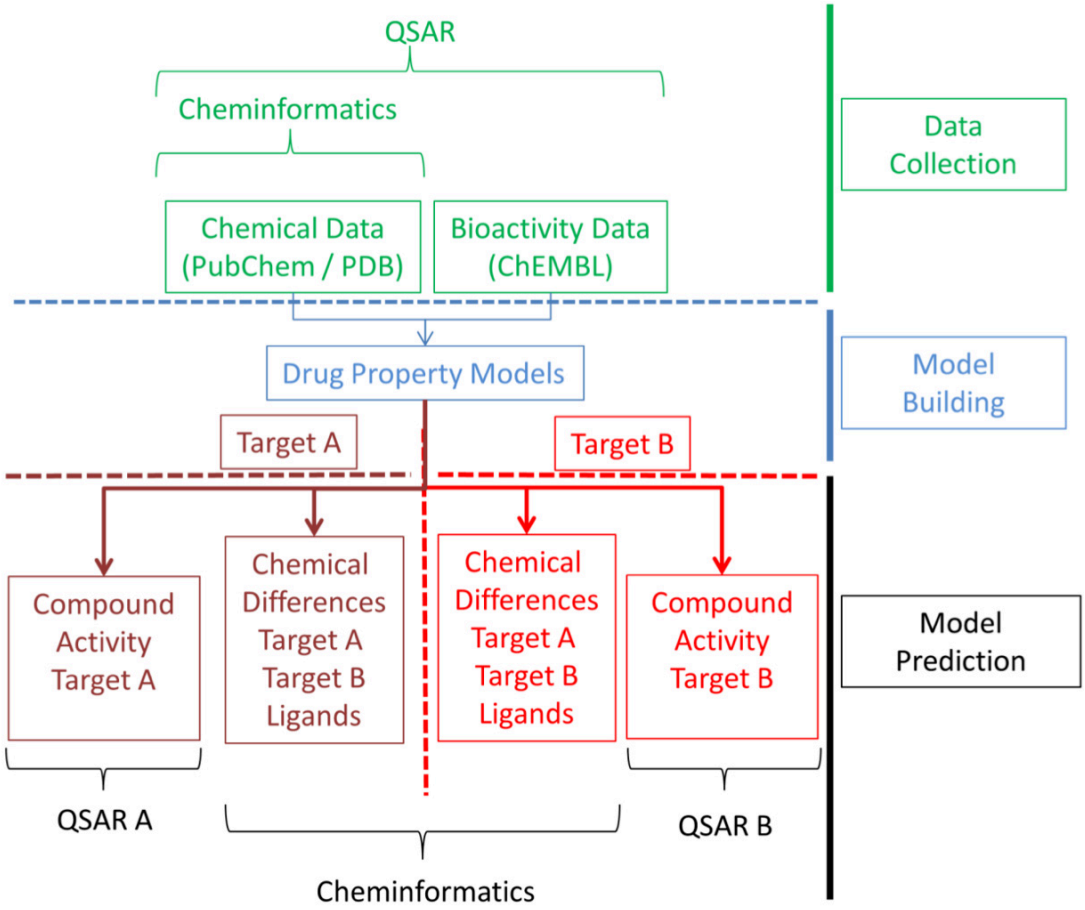


Figure 1.6: Simplified schematic overview of a cheminformatics project applied to a dataset consisting of two targets. The scheme also distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). Two statistical QSAR approaches to the same dataset are also shown. To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

1.7 Current Computational Bioactivity Modeling

1.7.1 No standardized tools. For bioactivity data, no general accepted method to process and mine large amounts of data currently exists. There are several methods to mine bioactivity data available, buts most are focused on small scale datasets. A rough distinction can be made between statistical methods like QSAR and more structural methods (see section **1.8** and **chapter 7**).

1.7.2 Quantitative Structure-Activity Relationships (QSAR). The term Quantitative Structure-Activity Relationship is often used to refer to one of two concepts. It can be a quantitative structure-activity relationship focused on a closely related ("congeneric") series of compounds, where the QSAR is driven by relatively slight variations in substitution patterns of the compounds. This is the classical QSAR concept. However, QSARs can also be statistical models that aim to explain the driving forces behind small molecules (ligands) interacting with a given protein (target). In this thesis, QSAR will mean the latter definition. As they are statistical models they do not provide a mechanistic explanation and rely heavily on machine learning. In this light, QSAR models are created ("trained") on a set of compounds for which the activity is known ("training set"). This chemical structure of the compounds in this training set is transformed into a way it can be processed by the computer (usually a numeric description of the compound called 'descriptors',[25] see **chapter 2** for further explanation). After a validation they can then be applied to a set of compounds for which the activity is not known ("test set") to discover novel compounds that display an activity on the target of interest.

Nevertheless, there are some limitations to this approach. Firstly, QSAR models are only able to make valid predictions for compounds that are at least similar to compounds in the training set (the so called applicability domain).[26, 27] This similarity is usually expressed based on the descriptor used to train the model and limited by the chemistry of compounds that have been previously tested on the target of interest. Furthermore, they can only make valid predictions for a single target since they are constructed on the chemical structures of a series of ligands and target information is not present. Finally, the quality of the QSAR (and hence the reliability of the predictions) is defined by the quality of the training set.

A simplified example application of a QSAR approach to a dataset consisting of two targets is shown in **Figure 1.6**. This dataset requires two separate QSAR models to be trained, one for each target. Each QSAR can subsequently predict the activity of ligands on that target. Data sources in this case can be databases like Pubchem or ChEMBL. The PDB can also be used, but needs to be combined with a database like ChEMBL to retrieve the actual bioactivity of known ligands.

1.7.3 Classification versus Regression. In machine learning an algorithm can be trained to predict one of two output variable types, each of which will be described below.

The first predicts a class as output variable (classification), in the simplest case the algorithm then decides whether or not the untested ligands belongs to either the 'active' class or the 'inactive' class based on the chemical resemblance to the training set of both the active compounds ('active' class) and the inactive compounds ('inactive' class). However, classification can also be performed using more than two classes and is limited only by computational power and memory.

Secondly, machine learning can predict a numeric output variable for an untested ligand (regression). In regression the predicted value (which can be any numeric value e.g. pK_i, pEC_{50}, pIC_{50}, etc.) is also calculated based on the similarity to the training set and the tested values for the output variable to be predicted. The advantage of regression over classification is that it provides a comparison between untested compound A and untested compound B. Once one of these untested compounds has a higher affinity it is presumed to be more active, whereas in classification both would be 'active'.

1.7.4 Validation of QSAR models. As outlined above, QSAR models can predict the activity of untested compounds on certain targets of interest. However, as they rely on statistics, these statistics have to be validated before any meaningful prediction can be made. Due to the different nature of classification and regression models, they require a different form of validation.

1.7.5 Classification Validation. In classification-based QSAR validation can be performed by analyzing the fraction of tested compounds that are classified correctly by the model rendering parameters like 'sensitivity', 'specificity', 'positive predictive value', negative predictive value'.[28] Each of these parameters outputs a value between 0 (poor prediction) and 1 (perfect prediction). In addition there is the Matthews correlation coefficient (MCC),[29] which aims to combine all four parameters in a single score between -1 (inverse prediction) and 1 (perfect prediction), in the case of the MCC being '0' also indicates a poor prediction (**Figure 1.7**).

To arrive at these numeric values, the predictions are divided into 4 types of predictions: True Positives (TP, compounds that are tested active and also predicted to be active), True Negatives (TN, compounds that are tested inactive and also predicted to be inactive), False Positives (FP, compounds that are tested inactive but predicted active) and False Negatives (FN, compounds that are tested active but predicted inactive).

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Figure 1.7: Definition of some of the validation parameters used to quantify the quality of classification-based QSAR models. TP stands for True Positives (compounds correctly predicted to be bioactive), TN stands for True Negatives (compounds correctly predicted to not be bioactive), FP stands for False Positives (compounds incorrectly predicted to be bioactive), and FN stands for False Negatives (compounds incorrectly predicted to not be bioactive).

1.7.6 ROC Plot. A number of machine learning algorithms, like Support Vector Machines (SVM), are capable of producing a ranking, indicating the likelihood that a predicted compound belongs to the class it is categorized in. This ranking can be used to create a Receiver Operator Characteristic (ROC) plot. In this plot the y-axis denotes the TP rate and the x-axis denotes the False Positive FP rate. To maximize the number of true positives a tradeoff can be made allowing additional false positives. When the ranked predictions are then plotted, the resulting plot shows the tradeoff for every possible threshold.[30] This curve provides a graphical interpretation of model performance (**Figure 1.8A**).

1.7.7 Regression Validation. In regression models, validation parameters can be calculated directly from the differences between the measured value and the predicted value for a certain compound as the model provides a numeric value. Usually a standard correlation coefficient (R^2) is calculated along with the Root Mean Squared Error (RMSE). However, most models are expected to predict a value for the output variable for a compound which is near identical to the measured value, an ideal model therefore predicts data points on a line that intersects the origin (0,0). Hence it is recommended to also calculate the R_0^2, the correlation coefficient while taking into account that the line should intersect the origin.[31]

In addition to calculating quantifiable values, the measured and modeled values for all compounds are also plotted in a scatter plot. This plot provides a direct intuitive overview of the model performance (**Figure 1.8B**); from the plot model bias (biased over- or under-prediction) becomes apparent.

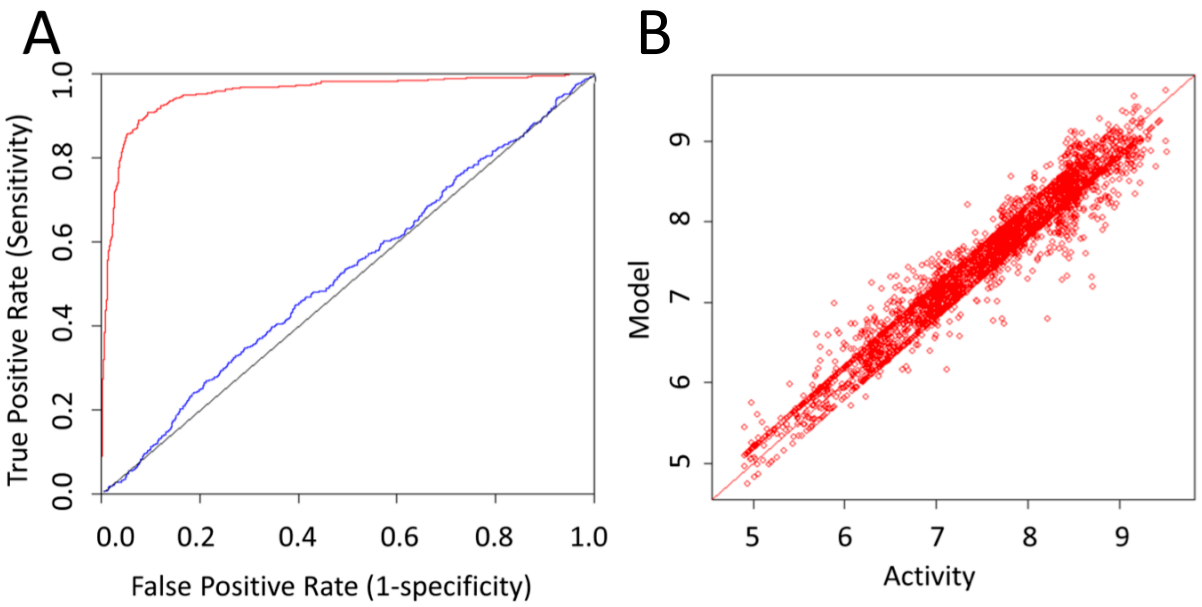


Figure 1.8: Validation plots in QSAR validation. (A) Classification validation using an ROC curve. The red line shows a highly predictive model reaching a high TP rate before trading off allowing more FPs. The blue line shows a poor model performing not much better than random showing a roughly equal TP and FP rate. (B) Regression validation using a measured versus predicted scatter plot. The scattering pattern of the plot already intuitively gives an impression of model performance. In addition R_0^2, R^2 and RMSE provide a quantifiable estimate for model performance.

1.8 Structural Methods

1.8.1 X-ray Crystallography. While X-ray crystallography is not a computational method itself, it is the most important data source for structure-based approaches and hence will also be introduced here. X-ray crystallography can be applied to both small molecules, entire proteins and proteins with a ligand bound. The technique uses the crystalline form of the analyte. Hence it is important that the analyte can be crystalized, which can be a major hurdle. Subsequently the analyte is subjected to a monochromatic beam of X-rays. These rays scatter as they travel through the analyte as they possess the correct wavelength to be scattered by the electron cloud of an atom (in the order of magnitude of Ångström, 10^{-10} m).[32] During this process the analyte is rotated.[32] From the angles and intensities of the scattered rays a crystallographer can produce the electron density map of the analyte. Subsequently, the mean position of atoms in the crystal and the bond orders between them can be determined via mapping onto this electron density map (**Figure 1.9**).

The high resolution information (down to a resolution of 1.4 Å) provides an excellent starting point for structure-based drug design. This is especially true when a known ligand is present in the crystal structure so that the part of the protein involved in the interaction ("binding pocket") is known from its orientation. Not only can this binding pocket be used to define protein similarity (**1.3.4**), complementary information about forces driving the interaction ("pharmacophoric information") can subsequently be obtained from both the binding pocket and the ligand. This information is the starting point for the design of novel ligands. However, a protein is not a static object; it is in fact very dynamic.[33, 34] Therefore a crystal structure can only be seen as a snapshot of one of the states a protein can exist in, but it does not provide any information about the number of other possible states the protein can exist in. However, this flexibility can be of great importance in drug design (See **chapter 7** for further details).[33, 35]

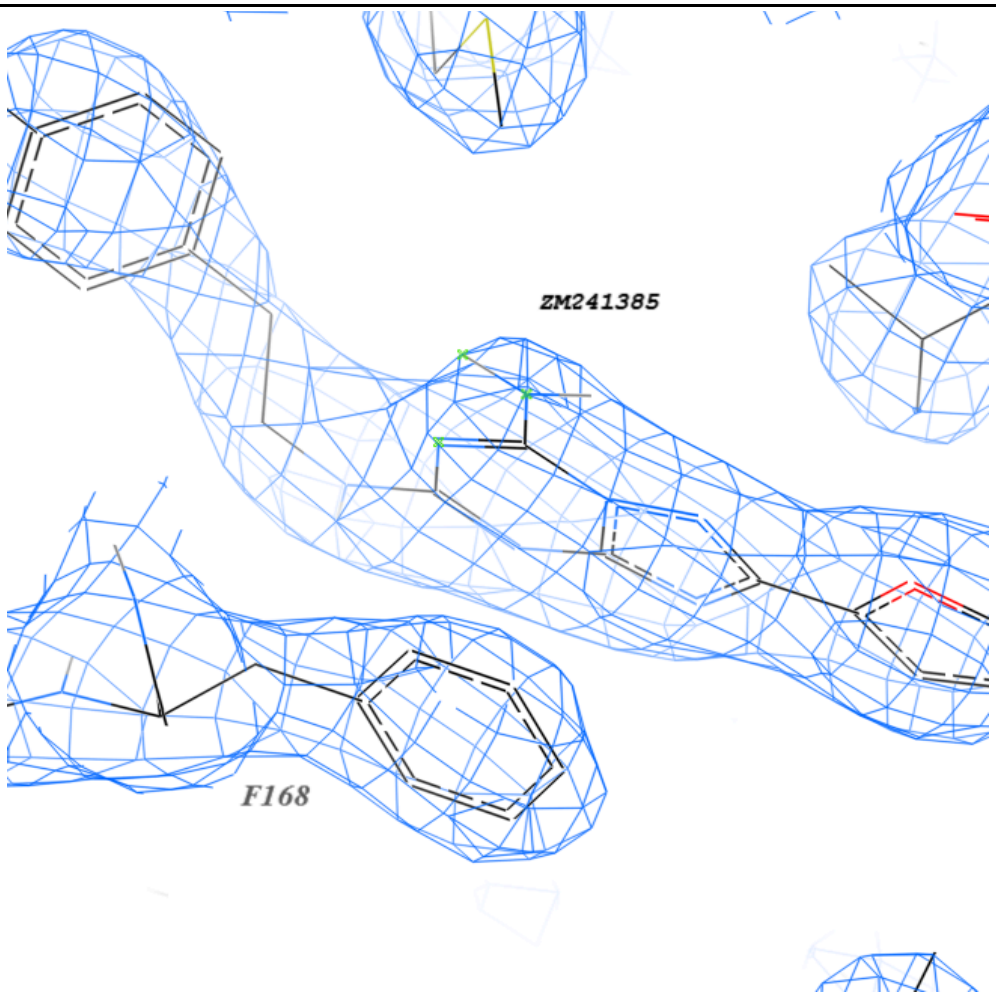


Figure 1.9: Electron density from PDB structure 3EML visualized as a grid in blue.[36] Shown within the density are the atoms that were mapped within this density. Both an amino acid from the protein (Phenylalanine, F168) and part of the ligand (ZM-241385) are visualized.

1.9 Combination of Computational Methods With 'wet' Experiments

1.9.1 Reliability issues. Computational methods have always been met with some reservations. While it is generally acknowledged that computational tools can present an unbiased view of the available data, it has also been shown that blindly following algorithms can lead to expensive experimental failures. However, over the last years the methods have become better accepted in existing research programs. One of the crucial factors to a good integration is reliable predictions, because if one cries wolf too often the computational chemist will lose all credibility. The key to prevent false positives, or rather to minimize the chances of FPs, is to perform a small scale validation of the computational approach that accurately simulates the way it will be integrated in existing research programs. This means that it is absolutely essential to prospectively validate at least some of the predictions made by any model using an actual wet experiment. It also means that is essential to disregard any model unless TPs are identified in such a prospective approach.

1.10 Informatics Approaches for Bioactivity Data

1.10.1 Proteochemometric modeling. As mentioned before, for bioactivity data, no generally accepted method to process large amounts of data exists. An illustration of the reasons why current methods are ill equipped to deal with large amounts of bioactivity data will be outlined below. An overview of the different computational methods mentioned before is given in **Figure 1.10**, again in the form of a scheme for a medicinal chemistry project involving two targets.

The cheminformatics approaches focus on chemical data; therefore these methods can make predictions about chemical properties relevant for different targets. Bioinformatics focusses on protein data; therefore these methods can make predictions about the differences between targets. QSAR links chemical data to bioactivity data and can therefore make predictions about the activity of compounds on a single target and also rationalize chemically why compounds are active.

Finally there is a relatively young method, proteochemometric (PCM) modeling,[37] which uses all three types of data. Hence it can make predictions about compound activity on multiple targets. Furthermore it can rationalize *why* a compound is active based on chemistry (features of small molecules) or based on biology (features of the proteins).

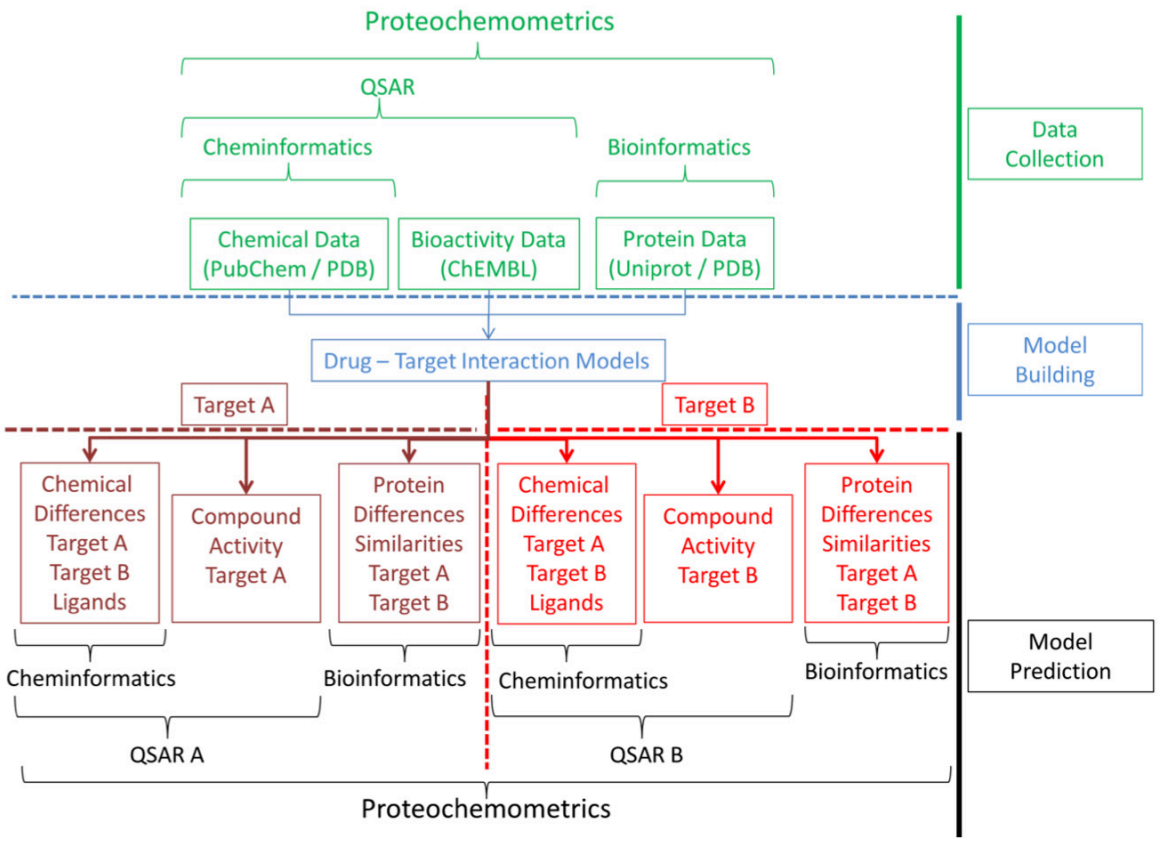


Figure 1.10: The different computational data analysis methods mentioned in this thesis, the data these methods process and their relationships to one another. This figure shows a schematic approach in the case of a medicinal chemistry project which involves two targets (deemed A and B). The scheme distinguishes three separate phases on the right hand side: data collection (green), model building (blue) and model prediction (black). To be able to distinguish between the different targets, they are visualized in different shades of red. See text for further details.

1.10.2 Statistical Methods. Statistical bioactivity modeling traditionally focusses around a single target. While some cheminformatics approaches have been introduced to combine several targets from a chemical point of view,[38] they do not possess the predictive abilities of classical QSAR approaches like affinity prediction and chemical interpretation of the SAR. Moreover, a single QSAR model is unable to explain selectivity e.g. in the case of the adenosine receptor subfamily

To explain selectivity using QSAR, individual QSAR models have to be constructed for each target. The differences between these QSAR models can then explain selectivity. However, this approach is already laborious for small groups of targets. Furthermore, the multiple QSAR approach cannot extrapolate to new (related) targets as it requires the creation of a novel QSAR model for that target. Hence, we cannot use this method to virtually identify small molecules that are active on a novel target as knowledge of compounds active on that target is required.

Therefore improvement of current methods is essential before bioactivity data can be processed on a large scale. An ideal approach would be able to use data gathered via bioinformatics (e.g. protein sequence) and combine that with chemical data (e.g. small molecules known to be active on a related target). PCM modeling might solve these shortcomings of QSAR as major bioactivity modeling approach. PCM is reviewed in **chapter 2** of this thesis.

1.10.3 Structural Methods. Structural methods are an advantageous tool to explain in detail what drives the interaction between a ligand and a target. However, due to the high level of detail they require a relatively strenuous interpretation. Furthermore, the interaction relies on both features from the ligand and from the target. Therefore a single structure is not always representative for all ligands that can bind a target. In addition, the aforementioned protein flexibility is difficult to extract from a single structure. Thus a need exists to partially automate the processing of structures, keeping unique features of interactions between a single ligand and target while at the same time also highlighting features that are shared between several ligands.

1.11 Aims of this thesis

In this thesis we want to investigate and benchmark new and recent techniques that are equipped to process bioactivity data on a large scale. These techniques should be able to link sets of related targets and ligands and therefore investigate the ligand – target space. As such PCM is a good (statistical modeling) candidate and it will be investigated in **chapters 2, 3, 4, 5**, and **6**.

Chapter 2 contains a literature review of PCM and other similar approaches carrying a different name. It highlights previous work, application areas and pitfalls. Subsequently **chapter 3** contains an extensive investigation of the physicochemical space of the natural amino acid side chains, and introduces four novel protein descriptors which we have developed and consequently applied in **chapters 4** and **5**. It also investigates co-variation between previously published amino acid descriptors and studies the ability of the descriptors to create bioactivity models.

Chapter 4 demonstrates a preclinical application of PCM, the goal being the discovery of novel hits (ligands) that are active on one or more targets (the adenosine receptors). **Chapter 5** highlights a later phase in drug discovery, namely the lead optimization stage. This chapter shows how PCM can be used to select the optimal candidate from a group of compounds that best inhibits a group of targets. Lastly, **chapter 6** focusses on a clinical application of PCM.

The technique is used to identify the optimal treatment regimen for individual patient based on the dominant viral genotype they are infected with. Summarizing, **chapters 4 - 6** cover several major phases in drug discovery, and the role PCM can play in that particular phase.

In order to also investigate more structural approaches we introduce in **chapter 7** 'Consensus Structures' and a technique to investigate the ligand – target space using crystal structures. Consensus structures are very similar to PCM as they rely on the combination of ligand information and target information, but differ as they rely on structural information rather than statistics.

Finally **chapter 8** contains general conclusions drawn from the thesis and future perspectives. In this chapter the focus is not so much on PCM but rather on computational methods in drug discovery in general.

Of the here mentioned approaches we will also characterize the limitations and possible pitfalls along with the ability to use these techniques on public data sets as they have become available. In this thesis we have been working both on G Protein-Coupled Receptors (GPCRs) and enzymes. Together these two classes are representative for most drug targets.

1.12 References

1. H. Meyer; *Zur theorie der alkoholnarcose.* Arch. Exp. Pathol. Pharmakol.; *1899*. **42**: 109-118.
2. E. Overton; *Studien über die narcose, zugleich ein beitrag zur allgemeinen pharmakologie.* Jena, Gustav Fisher *1901*. **45**: 195.
3. C.A. Lipinski, F. Lombardo, et al.; *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Adv. Drug Delivery Rev.; *2001*. **46** (1–3): 3-26.
4. P. Kirkpatrick and C. Ellis; *Chemical space.* Nature; *2004*. **432** (7019): 823-823.
5. C. Lipinski and A. Hopkins; *Navigating chemical space for biology and medicine.* Nature; *2004*. **432** (7019): 855-861.
6. E.E. Bolton, Y. Wang, et al.; *PubChem: Integrated Platform of Small Molecules and Biological Activities;* in *Annual Reports in Computational Chemistry*; A.W. Ralph and C.S. David; Editors. *2008*; Elsevier. p. 217-241.
7. E. Jain, A. Bairoch, et al.; *Infrastructure for the life sciences: design and implementation of the UniProt website.* BMC Bioinformatics; *2009*. **10** (1): 136-155.

8. H.M. Berman, J. Westbrook, et al.; *The Protein Data Bank* Nucleic Acids Res.; *2000*. **28**: 235-242.
9. A. Gaulton, L.J. Bellis, et al.; *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res.; *2011*. **40**: D1100 - D1107.
10. Intel Corporation. *Microprocessor Quick Reference Guide*. 2012 [cited 2012 January 5]; Available from: http://www.intel.com/pressroom/kits/quickreffam.htm.
11. Wikipedia contributors. *Instructions per second*. 2012 10 December 2011 [cited 2012 January 5]; Available from: http://en.wikipedia.org/wiki/Instructions_per_second.
12. J. Culver. *CPU Shack*. 2012 [cited 2012 January 19]; Available from: http://www.cpushack.com.
13. M. White; *Intel Historical Timeline.* Processor; *2003*. **25** (29): 9.
14. G. Shvet. *CPU World*. 2012 [cited 2012 January 18]; Available from: http://www.cpu-world.com.
15. D. Goodstein; *The Big Crunch*; in *NCAR 48 Symposium1994*: Portland.
16. M.F. Perutz; *Will biomedicine outgrow support?* Nature; *1999*. **399**: 299-301.
17. H. Moses, E.R. Dorsey, et al.; *Financial Anatomy of Biomedical Research.* JAMA: The Journal of the American Medical Association; *2005*. **294** (11): 1333-1342.
18. S.M. Sze; *Semiconductor devices: physics and technology*. 2nd ed.*2009*; New York: Wiley.
19. J.A.G. Briggs, T. Wilk, et al.; *Structural organization of authentic, mature HIV-1 virions and cores.* EMBO J.; *2003*. **22** (7): 1707-1715.
20. A. Touhami, M.H. Jericho, and T.J. Beveridge; *Atomic Force Microscopy of Cell Growth and Division in Staphylococcus aureus.* J. Bacteriol.; *2004*. **186** (11): 3286-3295.
21. G. Gulliver; *Observations on the sizes and shapes of the red corpuscles of the blood of vertebrates, with drawings of them to a uniform scale, and extended and revised tables of measurements.* Proceedings of the Zoological Society of London; *1875*: 474-495.
22. The UniProt Consortium; *Ongoing and future developments at the Universal Protein Resource.* Nucleic Acids Res.; *2011*. **39** (suppl 1): D214-D219.
23. W.M. David; *Bioinformatics: sequence and genome analysis*. *2004*; New York: Cold Spring Harbor Laboratory Press.
24. B. Frank K; *Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery;* in *Annu. Rep. Med. Chem.*; A.B. James; Editor *1998*; Academic Press. p. 375-384.
25. N. Nikolova and J. Jaworska; *Approaches to Measure Chemical Similarity – a Review.* QSAR Comb. Sci.; *2003*. **22** (9-10): 1006-1026.

26. H. Dragos, M. Gilles, and V. Alexandre; *Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models.* J. Chem. Inf. Model.; *2009*. **49** (7): 1762-1776.

27. L. Eriksson, J. Jaworska, et al.; *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environ. Health Perspect.; *2003*. **111** (10): 1361-1375.

28. P. Baldi, S. Brunak, et al.; *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics; *2000*. **16** (5): 412-424.

29. M. B.W; *Comparison of the predicted and observed secondary structure of T4 phage lysozyme.* Biochimica et Biophysica Acta (BBA) - Protein Structure; *1975*. **405** (2): 442-451.

30. J. Fogarty, R.S. Baker, and S.E. Hudson; *Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction*; in *Proceedings of Graphics Interface2005*; Canadian Human-Computer Communications Society: Victoria, British Columbia. 129-136.

31. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

32. B. Rupp; *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. 1st ed.*2009*; New York: Garland Science. 800.

33. H.A. Carlson; *Protein flexibility and drug design: how to hit a moving target.* Curr. Opin. Chem. Biol.; *2002*. **6** (4): 447-452.

34. H.A. Carlson and J.A. McCammon; *Accommodating Protein Flexibility in Computational Drug Design.* Mol. Pharmacol.; *2000*. **57** (2): 213-218.

35. K. Das, J.D. Bauman, et al.; *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: Strategic flexibility explains potency against resistance mutations.* Proc. Natl. Acad. Sci. U. S. A.; *2008*. **105** (5): 1466-1471.

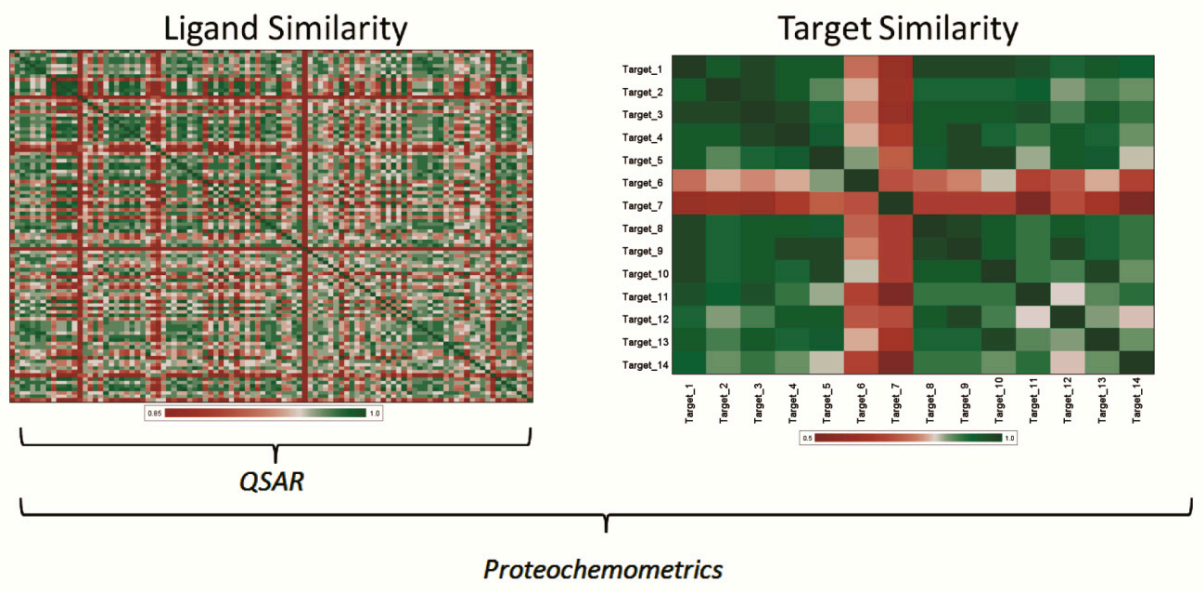
36. V.P. Jaakola, M.T. Griffith, et al.; *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science; *2008*. **322** (5905): 1211-1217.

37. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.

38. D.E. Gloriam, S.M. Foord, et al.; *Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design.* J. Med. Chem.; *2009*. **52** (14): 4429-4442.

Chapter 2

Proteochemometric Modeling as a Tool to Design Selective Compounds and Extrapolate to Novel Targets



G.J.P. Van Westen, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; Med. Chem. Commun.; 2011. ***2*** *(1): 16-30.*

Contents

2.1 Abstract

'Proteochemometric modeling' is a bioactivity modeling technique founded on the description of both small molecules (the ligands), and proteins (the targets). By combining those two elements of a ligand – target interaction, proteochemometric techniques model the interaction complex or the full ligand – target interaction space, and they are able to quantify the similarity between both ligands and targets simultaneously. Consequently, proteochemometric models or complex based models, can be considered and extension of QSAR models, which are ligand based. As proteochemometric models are able to incorporate target information they outperform conventional QSAR models when extrapolating from the activities of known ligands on known targets, to novel targets. Vice versa, proteochemometrics can be used to virtually screen for selective compounds that are solely active on a single member of a sub family of targets, as well as to select compounds with a desired bioactivity profile – a topic particularly relevant with concept such as 'ligand polypharmacology' in mind. Here we illustrate the concept of proteochemometrics and provide a review of relevant methodological publications in the field. We give an overview of the target families proteochemometric modeling has previously been applied to, and introduce some novel application areas of the modeling technique. We conclude that proteochemometrics is a promising technique in preclinical drug research that allows merging datasets that were previously considered separately, with the potential to extrapolate more reliably both in ligand as well as target space.

2.2 What is 'Proteochemometric Modeling' and what makes it useful for the design of bioactive compounds?

2.2.1 Structure-Activity Models. In 1962 Hansch *et al.* established the water/octanol partition coefficient (log P), discovered by Meyer and Overton,[1, 2] to quantitatively describe the relationship between the structure and biological activity of a substance using regression analysis;[3, 4] work that can be regarded as the first real Quantitative Structure- Activity Relationship (QSAR) study. Over the last decades this field has been greatly expanded, as computational methods can greatly reduce the number of experiments necessary to obtain a viable lead compound. They can do so by removing compounds from the set of candidates based on in silico experimentation before an actual 'wet lab' experiment needs to be performed.[5] The high expenses of innovative research and development have supported the case of computational research as it can be performed very cost effectively; i.e. it can help to reduce the costs of innovative drug research.[6] Furthermore the computational models obtained can be used to predict effects of untested substances, thus successfully finding their way into a virtual screening workflow.[7] Basic assumptions in structure activity studies are that (i) compounds sharing some chemical similarity should also share targets and (ii) targets sharing similar ligands should also share similar properties.[8-10] To summarize, conventional QSARs represent a very broad collection of computational tools that can model any output variable with input variables in the form of molecular descriptors using statistical approaches. However, QSARs have some limitations and drawbacks; these will be described in more detail below, followed by the extensions proteochemometric modeling makes to alleviate at least some of those limitations.

2.2.2 Why improve QSAR? A drawback of QSARs is that they consider the interaction of a group of compounds to only a single target, and often have a minimal ability to extrapolate (and sometimes even interpolate) into novel areas of chemical space.[11] This automatically requires that enough data is available on a specific target before a meaningful model can be constructed, which is rarely the case when searching for hits on a recently identified target. Furthermore, as conventional QSAR approaches consider only the ligands, the ability of QSAR approaches alone is very limited for identifying new classes of ligands or new binding modes of similar compounds outside the training set.

Although in practice there are usually multiple similar ligands that bind to a protein with varying affinities, these varying affinities are not caused only by the chemical structure, but also by the binding site. In fact, the concept of simultaneous considering ligand and binding site similarity has caused Kauvar to postulate that binding to any protein can be described by a linear combination of binding affinities to 'orthogonal' protein binding sites – a concept very much resembling current proteochemometric (and chemogenomics) thinking.[12]

The binding pocket is usually not a rigid pocket but has some flexibility present allowing an induced fit of the ligand molecule.[13, 14] As the pocket is not described by molecular descriptors in the case of QSAR, QSAR will naturally not always be able to describe all aspects of protein-ligand interaction.[15] Therefore, scientists should be cautious not to overstep the boundaries of the applicability domain for each QSAR (**Figure 2.1**). Overstepping this boundary might lead to the occurrence of 'activity cliffs' in the activity landscape present the modeler with situations where similar ligands do not always lead to similar activity.[13, 16-18] Possible causes of these cliffs include different binding modes, different binding sites and synergistic effects of chemical features of the ligand with features of the binding pocket; all of which cannot be covered in a QSAR model.

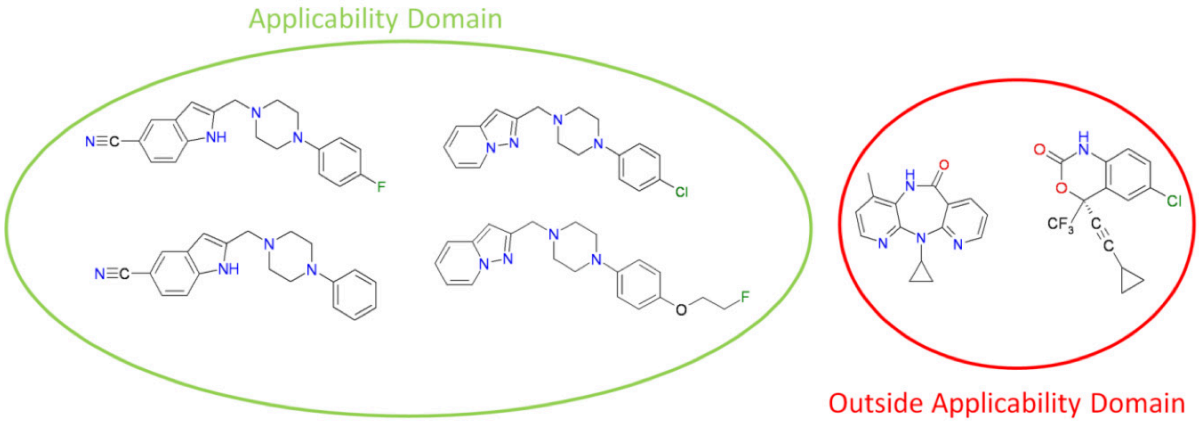


Figure 2.1: An example of the applicability domain concept. Depicted on the left side are known structures of Dopamine receptor binding compounds depicted. When a model is trained on these compounds it can only be expected to make reliable predictions for compounds that are chemically similar to this training set. On the right side two HIV reverse transcriptase inhibiting compounds are depicted, as these compounds are chemically very different from the training set the model cannot be expected to make reliable predictions of their possible affinity for the dopamine receptor.

2.2.3 Proteochemometric modeling. Contrary to QSAR, proteochemometric (PCM) modeling is based on the similarity of a group of ligands and a group of targets, to the extent that PCM models the so-called ligand-target interaction space.[14, 19] Like in QSAR modeling, the PCM model is constructed based on chemical descriptors that describe the compound data set and it introduces an additional term, a descriptor of the protein or target (**Figure 2.2**). Therefore a PCM model is constructed on both ligand and target similarity and can be regarded as an extension of conventional QSAR modeling. Furthermore one more additional term can be introduced, which describes effects on both ligand and target and the specific interactions between a compound and a target, called the cross term.[19-22] Here the difference with chemogenomic approaches becomes clear; chemogenomics is founded mainly on ligand similarities rather than the combination of the two. Nonetheless, the two techniques are quite similar and even show overlap as reviewed recently.[23] In a direct comparison Lapinsh *et al.* showed PCM to outperform QSAR and these findings were corroborated by both Geppert *et al.* and Ning *et al.*[19, 24, 25]

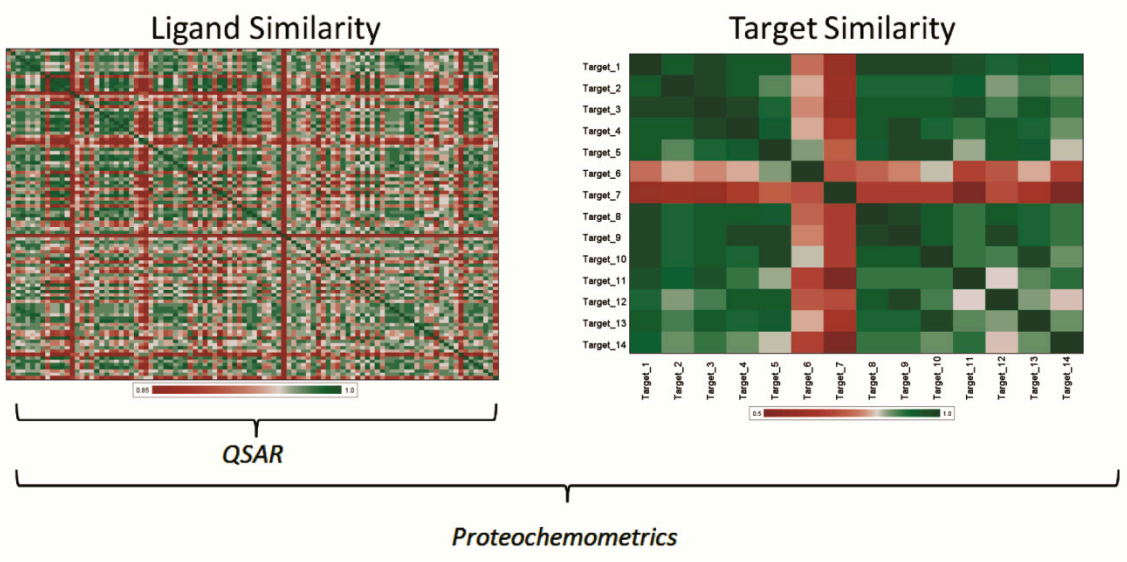


Figure 2.2: The difference between QSAR and PCM. The illustration shows two similarity matrices, one for a group of ligands and one for a group of related targets. In the heat maps green illustrates a high similarity while red illustrates a low similarity and white depicts an average similarity. PCM uses both ligand and target similarities for model generation, modeling the interaction complex. However, QSAR only uses ligand similarity, modeling only the left hand side of the ligand – target interaction space. Therefore, while QSAR and PCM are founded on similar principles, PCM can benefit from additional information in model training.

However, QSAR and PCM have also been shown to perform nearly identically on an identical training set by Lapinsh *et al.*,[11] although it should be noted that in that particular study a highly simplified form of protein description was used. The main advantage of PCM is that the model can describe different interactions of a series of compounds to a series of targets while still being able to describe specific interactions of individual compounds to individual targets in the data set. Effectively PCM can thereby connect neighboring QSAR datasets on the basis of the similarity between the targets contained in these data sets. Therefore, in order to create a true PCM model, it is necessary to have activity data of multiple compounds on multiple targets. When creating a PCM model on a single target, the fact that all targets are identical makes this PCM model in reality a QSAR model.[21, 26]

The fact that PCM can connect neighboring QSAR datasets makes it quite similar to inductive learning. Inductive learning is a QSAR based technique consisting of the learning of a property on a dataset and to use the extracted knowledge to better learn a related property (e.g. learn rat blood-brain barrier permeability based on a large rat dataset, then use the predicted rat BBB term as a descriptor in a human BBB model. Effectively only the difference between human and rat should then be learned, while letting the rat QSAR model account for the common issues that modulate BBB-permeation in both species). In the example given here, the predicted rat BBB descriptor can be compared with a protein descriptor in the case of PCM. The major difference is between the example and PCM is that the former requires two separate steps of model training where the latter requires one. Furthermore the interpretability of PCM will likely be higher as it relates to target (dis)similarity rather than a non-target related descriptor such as 'rat BBB penetration'

Since PCM contains a target descriptor, the major advantage is that PCM can create a single model predicting a single output variable of the interaction, e.g. affinity, between a very diverse series of compounds or targets and still provide a statistically solid model.[11, 20] It allows the modeler not only to extrapolate the activity of new compounds on known mutants or targets, but also to extrapolate the activity of known compounds on new mutants or targets (**Figure 2.3**). PCM can also be applied in situations where the 3D information of the targets is unavailable or when the 3D approach is unreliable. Typical examples are situations where no crystal structure is at hand or where only low quality homology models are available.

PCM, like QSAR, can implement a variety of machine learning techniques including both linear and non-linear methods to construct a model.[19, 27] Furthermore PCM has already been applied to a wide variety of relevant drug targets. To illustrate this versatility we will provide a short overview of the targets PCM has previously been applied to below. Subsequently we will provide an overview of previously used ligand descriptors, target descriptors and cross terms in PCM modeling. Thirdly we will outline some of the machine learning techniques compatible with PCM. We will end the review with some possible pitfalls and disadvantages and the final conclusions.

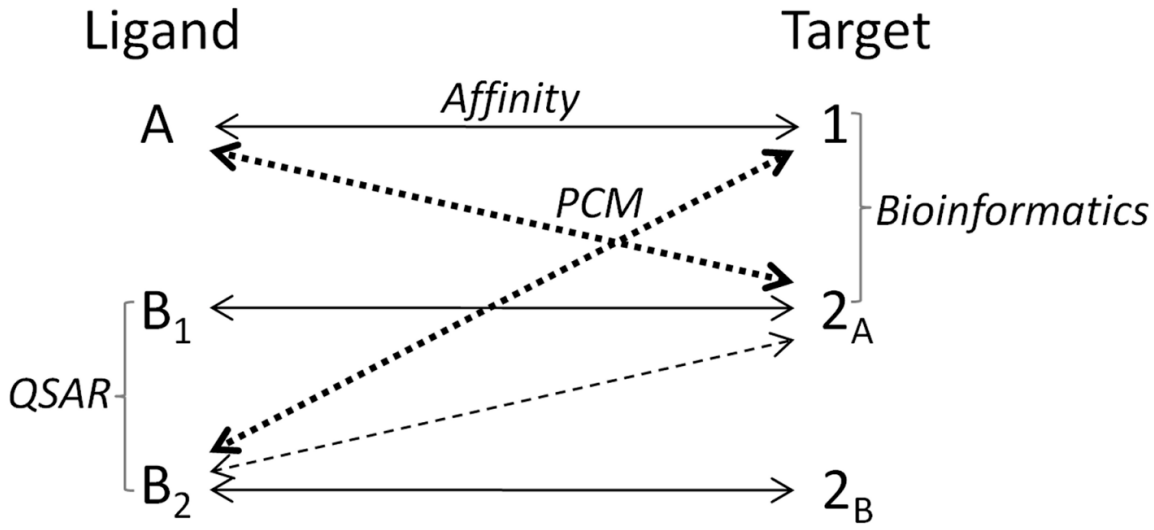


Figure 2.3: Possibilities of PCM in a hypothetical dataset where the affinity of three different compounds was measured on three different targets. QSAR is able to calculate an output variable based on the similarity of compounds (use the similarity between compound B_1 and B_2) and Bioinformatics can quantify the similarity between targets (similarity between target 1 and 2_A). However PCM can use this information to extrapolate the activity of compounds on targets (dashed double arrows). In this case a PCM prediction for the activity of B_2 on 2_A is likely more accurate than the prediction of the activity of B_2 on target 1.

2.3 Biochemical applications of PCM techniques

2.3.1 G Protein-Coupled Receptors. Although PCM has been introduced fairly recently the technique has been tested on a wide variety of relevant drug targets. For a comprehensive overview of targets to which PCM has been applied see **Table 2.1**. Overall PCM has mainly been applied to datasets of G protein-coupled receptors (GPCRs), in particular the rhodopsin-like class A receptors. Dopamine, histamine, adrenergic and melanocortin receptors have been described in various ways ranging from binary to local descriptors of protein structures.[28] Interestingly, Lapinsh *et al.* showed that PCM was able to create a viable model from a dataset containing multiple related class A receptors based on their transmembrane alpha-helical regions with a pKi error of approximately 0.55 log units, which is a rather well-performing model.[14] It should be noted though that this dataset contained limited chemical diversity as it was based on only 22 ligands. The targets side contained 31 GPCRs, being described by 159 TM domain amino acids. In related work, Weil and Rognan also create a model including multiple class A GPCRs. Their model was based on a custom fingerprint encoding both ligand and target features in one fixed length array of bits.[29] Using several classification models, they showed that it is in fact possible to create one global GPCR PCM model. Furthermore, they demonstrated that their models are able to retrieve the natural receptor ligands from a decoy-spiked dataset of 200,000 ligand – target pairs.

Likewise, Bock *et al.* modeled multiple GPCR receptor families in a single model, a PCM based approach, which they applied to orphan GPCRs.[30] Using a Support Vector Machines (SVM) learning approach they were able to separate a small group (2%) of highly active ligand – compound pairs from the bulk of their 1.9 million data point dataset. As an extension of this work, Jacob *et al.* applied a PCM approach to the GLIDA GPCR database to predict the ligands of orphan GPCRs.[31] They achieved a prediction accuracy of approximately 90 %.

Table 2.1: List of applications of PCM modeling (near-comprehensive to the knowledge of the

Year	Modeling Technique	Targets	Validated
2001	PLS	Melanocortin Receptors	No
2001	PLS	1a, 1b and 1d Alfa-Adrenergic Receptors	No
2002	PLS	Serotonin, Dopamine, Histamine, and Adrenergic receptors	No
2002	SVM	CLiBE Selection	No
2003	PLS	Melanocortin Receptors	No
2005	PLS	Melanocortin Receptors	No
2005	PLS	Set I Serotonin, Dopamine, Histamine, Adrenergic receptors Set II 1a, 1b and 1d Alfa- Adrenergic Receptors	No
2005	PLS	Serotonin, Dopamine, Histamine, Muscarinic Acetylcholine and Adrenergic receptors	No
2005	SVM	Orphan GPCRs	No
2006	PLS	Melanocortin Receptors	Yes
2006	RS	Set I Melanocortin Receptors Set II Melanocortin Receptors Set III Adrenergic Receptors	No
2006	RS and PLS	Hydrolases, Lysases, Neuramidases, Anhydrases	No
2006	PLS	PDBind subset	No
2007	PLS	Melanocortin Receptors	No
2007	PLS	Antigen recognizing Antibodies	No
2007	PLS	Melanocortin Receptors	No
2008	PLS	(point mutated) HIV Proteases	No
2008	PLS	Cytochrome P450 enzymes	No
2008	Linear and NN	Matrix Metalloproteinases	No
2008	PLS	Dengue Virus NS3 Proteases	No
2008	SVM	Large Crystal Structure data set	No
2008	SVM	GLIDA subset	No
2009	SVM	Proteases	No
2009	PLS	(point mutated) HIV Proteases	No
2009	PLS	(point mutated) HIV Proteases	Yes
2009	SVM, RF, NB	MDL Drug Data Report subset	No
2009	SVM	117 Pubchem Targets	No
2009	SVM	DrugBank	Yes
2010	PLS	Major Histocompatibility Complex Proteins	No
2010	SVM, DT, NB	Multi target BindingDB dataset	No
2010	DT, NN, SVM, PLS	Kinase Inhibitors	No
2010	SVM	Kinase inhibitors (ProLINT database)	No

Explanation of abbreviations: Protein Database (PDB), Computed Ligand Binding Energy (CLiBE), GPCR LIgand Database (GLIDA), Random Forest (RF), Naïve Bayesian (NB), Decision Tree (DT), Grid Independent Descriptors (GRINDs),

authors, though publications using a different name for the technology might have been missed):

Ligand Descriptors	Target Descriptors	Cross Terms	Ref.
Binary	Binary	Multiplication	28
Binary	Sequential (Z-scales)	Multiplication	19
GRINDs	Sequential (Z-scales)	Multiplication	62
1D projection of Physicochemical properties	Sequential Physicochemical Properties	-	42
GRINDs	TM Identity	Multiplication	11
Physicochemical	Binary	Multiplication	20
GRINDs	Sequential (Z-scales)	Multiplication	85
GRINDs	Sequential (Z-scales)	Multiplication	14
Physicochemical and 2D	Sequential Physicochemical Properties	-	30
Binary	Sequential (Z-scales)	Multiplication	21
ASCII String	ASCII String	-	27
Physicochemical, 1D, 2D, 3D	Local Descriptors of Protein Structure	Multiplication	35
2D and 3D	3D Structural	Protein-Ligand Interaction	67
Binary	Binary / Sequential (Z-scales)	Multiplication	65
Sequential (Z-scales)	Sequential (Z-scales)	Multiplication	36
Sequential (Z-scales)	Sequential (Z-scales)	Multiplication	26
GRINDs	Sequential (Z-scales)	Multiplication	32
GRINDs	Binary and Sequential	Multiplication	41
2D autocorrelation vectors	Amino Acid Sequence Autocorrelation vectors	-	93
Physicochemical	Binary	Multiplication	34
Physicochemical, 1D, 2D, 3D	Local Descriptors of Protein Structure	-	70
2D and 3D kernels	Multitask, Hierachy and Binding pocket kernel	-	31
2D Fingerprints	Sequence Similarity	-	24
Physicochemical	Sequential (Z-scales)	Multiplication	33
Sequential (Z-scales)	Sequential (Z-scales)	Multiplication	112
2D, Shannon Entropy, Pharmacophoric	Pysicochemical property based phylogenetic tree	Merger of ligand and target descriptors	29
Topological Graph Based	Sequence Similarity	-	25
2D graph vector	Physicochemical property vector	-	113
Sequential (Z-scales)	Sequential (Z-scales)	Multiplication	40
Physicochemical	Sequence Similarity (multiple descriptors)	-	46
Physicochemical, Geometrical, Molecular	Sequential (Z-scales) + Sequence Similarity (multiple descriptors)	Multiplication	39
2D autocorrelation vectors	Amino Acid Sequence Autocorrelation vectors	-	38

one dimensional (1D), two dimensional (2D), three dimensional (3D), Human Immunodeficiency Virus (HIV), Partial Least Squares (PLS), Rough Set (RS), Neural Net (NN), Support Vector Machines (SVM), Reference (Ref.).

2.3.2 Viral Targets. While GPCRs are quite amenable to PCM modeling due to their relatedness, the same is true for mutants of enzymes. Hence, another target type that has been adequately covered in PCM modeling are viral proteins, such as HIV Protease,[32, 33] Dengue virus NS3 Protease,[34] and Influenza virus A and B Neuraminidase.[35] The average similarity between these targets is very high when compared to the average similarity between multiple GPCR families. This is especially true in the case of HIV proteases, where the differences between targets may be a single amino acid.

2.3.3 Other macromolecules. In addition to these larger target groups, PCM has also been applied to antigen recognizing antibodies,[36] matrix metalloproteinases,[37] kinases,[38, 39] Major Histocompatibility Complex (MHC) proteins,[40] and Cytochrome P450 enzymes.[41] The ligands PCM has been applied to include both small molecules and peptides. Furthermore, the output variables modeled by PCM models include classification,[29, 31] binding affinity,[20] and even equilibrium binding free energy.[42] The nature of the datasets PCM has been applied to confirm the potential of PCM as a versatile technique suitable for any dataset. Whether the targets are highly related or more dissimilar, a functional model that provides better extrapolation capabilities than QSAR can be created. However, it should be noted that the type of target description must be tuned to the nature of the dataset.

2.4 Novel applications of PCM

2.4.1 Hit identification for orphan targets. The main advantage of the PCM extension of conventional QSAR modeling is that it allows the merging of datasets that describe the affinity to highly similar but not identical targets; hence it allows for the extrapolation of affinity modeling between these datasets. Thereby PCM can fulfill a need in hit identification for newly identified targets – including applications such as the prediction of polypharmacology or the deorphanization of receptors. PCM not only models similar datasets simultaneously, it also quantifies the distance between the different targets. Therefore it allows the scientist to estimate the reliability of any model prediction depending on the distance of both the ligand and target to the training set, through determining the applicability domain of the model.

2.4.2 Simultaneous modeling of orthosteric and allosteric ligands. PCM includes a target description in addition to the ligand descriptors; hence it is able to quantify the similarity between different binding sites. As a result it could be used for the simultaneous modeling of a series of orthosteric and allosteric ligands of a single target even though they act through a different binding site.[43-45] In this case not two (or more) proteins are used for modeling the target side, but rather two binding sites of ligands that are in different parts of the protein which are able to accommodate completely different ligand chemistry. Previously it has already been shown that targets sharing a low similarity and accommodating a completely different chemistry can be modeled successfully with PCM.[35, 46]

Capturing the chemical information of different types of ligands that act on different binding sites on a single target can provide advantages as the increase in information that serves as model input could lead to better extrapolation capabilities of a single target model. The single target model can in this case be seen as a two targets model. (**Figure 2.4**)

A second possible application is the simultaneous modeling of a drug regimen incorporating both nucleoside reverse transcriptase inhibitors (orthosteric) and non-nucleoside reverse transcriptase inhibitors (allosteric) for a single dominant HIV mutant in a patient (**Figure 2.4**).

Finally, combining allosteric and orthosteric information in one model can provide advantages in preclinical research when looking for novel allosteric inhibitors or enhancers of proteins. Allosteric drugs have been shown to provide advantages in treatment by better resembling physiological signaling.[47] Furthermore these models could come in use, when research is targeting a protein for which the orthosteric (natural) ligand is known and part of an essential physiological process. As this process cannot be completely disrupted (orthosteric inhibition) or continuously activated (orthosteric activation or agonism) the possibility of allosteric modulation is very promising here. The addition of the orthosteric inhibitors to this model potentially could potentially detect cross reactivity with the orthosteric binding site at an early stage.

While to our knowledge this research has not been performed yet, it illustrates the versatility of PCM modeling approaches. PCM can cover a much larger bioactivity space than conventional QSAR models alone as well as, introduced here, multiple modes of action.

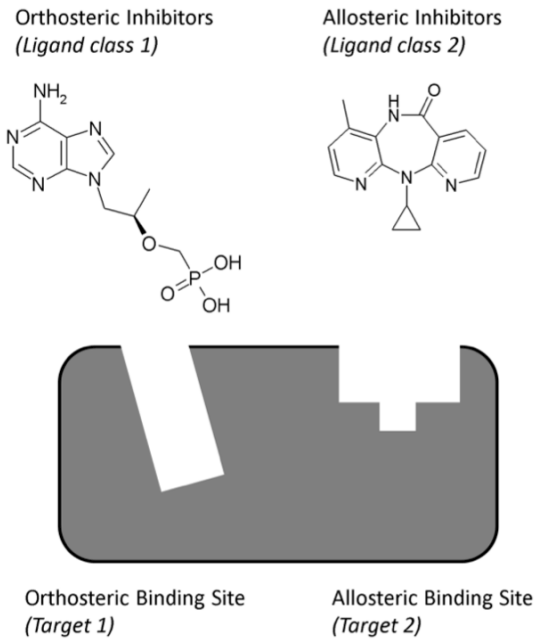


Figure 2.4: A single PCM could also potentially be used to model both allosteric and orthosteric binders to a given target, as an extension of current PCM models. Hereby a single target model is effectively turned into a multi target model, with the additional variable being introduced the binding site a given ligand binds to, this being particularly relevant in cases where a single chemical class of molecules can bind to each of the receptor subsites. Shown here are Tenofovir (an orthosteric HIV Reverse Transcriptase inhibitor) and Nevirapine (an allosteric HIV Reverse Transcriptase inhibitor). While modeling of this type has not been performed yet and its precise performance remains to be validated in the future, it still illustrates the ability of PCM models to incorporate not only ligand and target variables, but also additional variable types such as those of subpockets of a given protein target.

2.5. Ligand descriptors

Ligand descriptors are merely numerical methods to describe either properties of or differences between the compounds to be modeled, therefore a large number of different descriptor methods for ligands are available (For reviews see [8, 10, 13, 48, 49]). In this review article we will restrict our focus on ligand descriptors previously used in PCM. There is no optimal descriptor for all datasets and it is therefore wise to sample several different descriptors to identify the optimal descriptor for each setting.[13] In the following we will start our discussion with the simplest descriptors used in PCM, namely binary descriptors.

2.5.1 Binary compound descriptors. Binary descriptors are descriptors based on one or more binary flags set to 1 or 0. These descriptors can be based on the differences between functional groups on one common scaffold. Lapinsh *et al.* implemented them in PCM by creating a table listing all possible combinations of three functional groups and assigning unique binary descriptors to each of those combinations.[19] These descriptors are relatively simple and therefore computationally very fast; however results from the model have to be translated back to the functional group combination before model results can be interpreted.

Furthermore predicting values outside the range of descriptors is not self-evident, so there is little possibility of numerical inter- and extrapolation.[23, 50] Therefore, the authors recommend not to use binary descriptors in PCM. Additionally, binary descriptors function on the assumption that it is already known which properties are relevant in the QSAR. Furthermore they are very sensitive to the creation of 'ties',[51] namely the creation of equidistant similarity distances for non equal compound pairs. Binary descriptors are also more affected by possible overfitting of the data, since they do not allow a fuzzy interpolation of relevant groups.[52] One-Dimensional (1D) numerical (real-valued) descriptors resolve most of these problems and are better interpretable.

2.5.2 One dimensional and physicochemical compound descriptors. The main advantage of one-dimensional descriptors is that they are quickly calculated and modeled. Physicochemical descriptors are a subtype of these 1D descriptors, such as molecular weight, polar surface area or polarizability. When compared to binary descriptors, the usage of physicochemical descriptors increases the interpretability of the model and makes a translation step obsolete. The downside of this subtype of descriptors is the high number of possible descriptors that can be calculated for every compound. Therefore it is common to make a selection of descriptors that are important in each interaction model while also removing possible covariance in the descriptors. This goal can be achieved by preprocessing of the descriptors before a reliable model can be constructed, leading to the risk that only those descriptors are selected that are applicable to the training set on which the model is constructed. 1D descriptors have previously been applied successfully in PCM allowing the creation of a predictive and highly interpretable model (reaching an R^2 of 0.92).[20]

2.5.3 Two dimensional topological compound descriptors. Topological descriptors are two dimensional representations of compounds, visualizing bond properties, atomic properties and the inner atomic distances between the functional groups. These descriptors can be transformed into molecular graphs, a method widely used in substructure searching and clustering. In graph descriptors, the atoms form the nodes of the molecular graph and the bonds the edges.[10] Advantages of this type of descriptors are that they are relatively quickly computed and easily interpreted, but a downside can be that they lack three-dimensional (3D) information (for a recent review see Van der Horst *et al.* [53]). To the knowledge of the authors these descriptors have only been used in a PCM-like approach by Ning *et al.*[25] Their particular graph-based descriptor performs quite well and has previously been shown to perform comparably to 2D circular fingerprints, which are mentioned below.[50]

2.5.4 Two dimensional circular fingerprints. Descriptors based on 2D fingerprints convert the two dimensional information of the compound structure into a linear binary string. This class of descriptors is overall similar to 2D graph-based approaches, however it tends to be computationally much faster.[10] These 2D fingerprints are constructed from several substructures present in the compounds while the substructures are limited by a radius defined by a certain number of covalent bonds.[54, 55]

Circular fingerprints have previously been found to capture a large amount of information and to provide a high retrieval rate while at the same time being chemically interpretable as individual (localized) features.[54, 56] In recent studies the authors have been applying circular fingerprints in PCM models of HIV Reverse Transcriptase producing models with an average prediction error of 0.5 log units.[57, 58]

2.5.5 Alignment based 3D compound descriptors. 3D descriptors preserve more information of the compounds by adding information concerning the conformation of the compound.[59] However, most 3D descriptors require superposition of the compounds in their active conformation in 3D space. Only then useful information can be obtained, making the calculation more complex. The process of compound superposition is error prone and can introduce more noise than functional information,[60] a step which can be avoided by using internal distances between atoms or surface points of a compound. This translation preserves all possible pharmacophore triplets and quadruplets, features of the compound and inter-feature distances but simplifies the calculation step.[10] The increase of information in these descriptors also increases their size and consequently the calculation times of the models. To the knowledge of the authors, 3D descriptors have as yet not been used in PCM and the previously mentioned disadvantages might support the usage of GRINDs instead (see next section).

2.5.6 Grid independent descriptors. Grid Independent Descriptors (GRINDs) are descriptors that are obtained starting from a set of molecular interaction fields using different probes. This procedure involves a first step, simplifying the fields, and a second step, encoding the fields into alignment-independent variables using an autocorrelation transform.[61] The obtained descriptors can also be used to provide graphical diagrams and the original descriptors can be regenerated from the transformation in order to visualize the results of the analysis graphically in 3D. As one of few 3D descriptors GRINDs have previously been used in PCM with a prediction error of around 0.5 log units on datasets of GPCRs.[11, 62] This confirms the compatibility of PCM with GRINDs on this dataset. In this case, the GRIND descriptors were preprocessed using principal component analysis (PCA), a step to handle the high dimensionality common to many descriptor spaces.[11]

2.6 Protein descriptors

The main difference between ligand and protein descriptors is that the protein is, in general, a larger structure to describe and hence in most cases a selection of a subset of the residues (such as those lining the binding site) would be recommended. This selection can be made on the basis of a crystal structure if available, on data from mutational experiments or from an information-based bioinformatics analysis, e.g. a two-entropy analysis.[63]

At the level of the residues, a distinction can be made between descriptors that describe a (sub)structure or general property of the protein and those that stand for properties of individual amino acids on a sequential basis. Protein descriptors have been reviewed extensively and only methods that have been previously used in PCM will be highlighted here.[10, 64]

2.6.1 Binary protein descriptors. Similar to binary ligand descriptors, binary description of proteins can be performed based on several binary flags corresponding to different substructures of the protein. Although the make-up of the data set dictates the binary method used for the description and hence the length of the descriptors, binary descriptors are generally fast from a computational point of view.

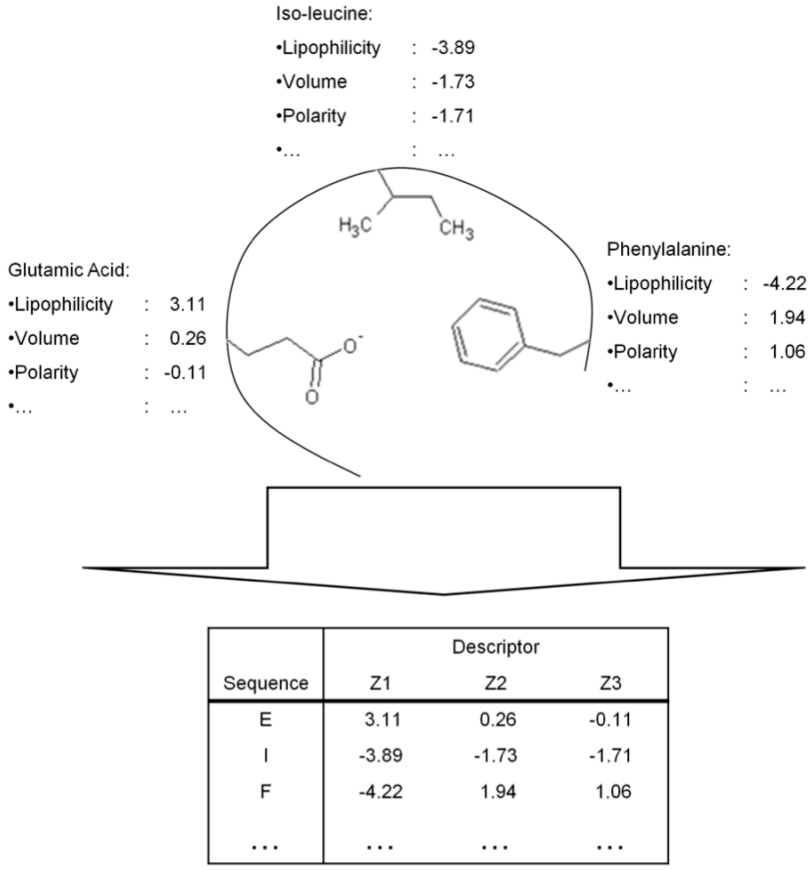
An example of a binary descriptor in PCM is given by Kontijevskis *et al.* who created several chimeric proteins divided into five segments based on building blocks obtained from four different receptors.[65] Thereby every segment was described by four binary descriptors and every protein by five segments, leading to each protein being described by 20 binary descriptors and to a unique descriptor for every protein. When directly compared with a sequential protein descriptor the binary descriptors (Root Mean Squared Error (RMSE) 0.61) outperformed the sequential descriptors (RMSE 0.76). However, in this form it is far less interpretable than sequential descriptors and limited in extrapolation capabilities.

A related subtype of binary protein descriptors is a feature-based semi-binary protein descriptor. In this descriptor each individual amino acid gets assigned a unique identifier resulting in a unique descriptor for each sequence.[57, 58] The final model is then trained on the collection of unique identifiers per sequence. The authors found this type of descriptors to outperform sequential descriptors, as is the case with the previously mentioned binary descriptors. The main advantage of these descriptors is that they are better interpretable than the binary descriptors as important residues can be individually identified rather than identifying an entire important subsection of a protein.

2.6.2 Three dimensional protein descriptors. While the descriptors used above encode only 2D properties of protein sequences, the targets are three-dimensional entities and this information can also be used in PCM modeling. Considering multiple mutants of HIV Reverse Transcriptase in parallel, Van Westen *et al.* used protein energy fields as descriptors for the investigation of complexes between mutant forms of HIV-RT and different ligands.[66] This appeared very useful in understanding the molecular mechanism and in suggesting novel chemistry ideas for anti-resistant inhibitor design. A 3D protein descriptor taking into account Cα coordinates and ϕ/ψ angles is introduced by Lindström *et al.*[67] Alignment problems and the large number of variables emerging from these descriptors were circumvented by preprocessing the data by PCA and covariance transformations. Lindström et al. show that this form of descriptor contains sufficient information to generate both global and sub-class specific PCM models.[67]

Contrary to full protein 3D descriptors, an example of a (local) 3D protein descriptor was introduced by Hvidsten and Kryshtafovych.[68, 69] Here 3D substructures are a collection of continuous short backbone fragments centered on a single residue within a predefined radius. These descriptors provide a highly generalized descriptor of the protein binding pocket and can therefore be used between completely different classes of proteins without the need for a multiple sequence alignment. They have been applied to PCM modeling of a very diverse target library of PDB structures by Strombersson et al.[35, 70] Performance was limited (cross validation prediction error of 1.7 log units), although it should be taken into account that this was a very diverse data set. We conclude that these descriptors can be especially useful when building a PCM model of a diverse dataset for which 3D structures are available.

2.6.3 Sequential protein descriptors. As for nearly all possible drug targets (human proteins) the amino acid sequence is available, one can use information derived from this sequence as a protein descriptor. Jacob *et al* employed this information to create several similarity measures.[31] In this case a hierarchical tree of the sequences guided a division in classes and through these classes a similarity measure of the binding pockets is obtained. They showed that using sequence as a protein descriptor enables the creation of a PCM model, obtaining a prediction accuracy of 90 %.[31] This performance can most likely be improved by converting a purely sequence based protein descriptor to a sequential descriptor of physicochemical properties of the amino acids, as described below.



	Descriptor		
Sequence	Z1	Z2	Z3
E	3.11	0.26	-0.11
I	-3.89	-1.73	-1.71
F	-4.22	1.94	1.06
…	…	…	…

Figure 2.5: Conversion of physicochemical properties of amino acids in the binding site into a protein descriptor via sequential protein descriptors using the Z-scales as an example. Based on the physicochemical properties of the amino acids side chains a protein descriptor is constructed. A trained PCM model can subsequently compare the binding sites of the proteins based on the differences in the physicochemical properties of the side chain.

Amino acids can also be described according to their physicochemical properties, like descriptors of small molecules. Several descriptor scales are available from the field of peptide drug modeling.[71-74] In PCM, the 'z-scales' introduced by Sandberg et al. that describe the properties of the amino acids, have been shown to perform well in many cases (**Figure 2.5**).[19, 67] The z-scales are based on a PCA used to compress a large input matrix that describes a broad selection of physicochemical properties into 5 principal components (PCs). In a plot of PC1 versus PC2, similar amino acids are located close together and dissimilar amino acids are spread further apart (**Figure 2.6**). PC1 can be interpreted as a lipophilicity scale and PC2 as volume/polarizability scale.[73] PC3 mainly describes polarity while PC4 and PC5 are more difficult to interpret. It has been shown that these sequential descriptors can predict contributions to selectivity without 3D information available 20 or predict the site of origin of indirect effects on ligand recognition by proteins.[21]

One possible problem with the z-scales or similar descriptors is the slightly limited interpretability since these descriptors are obtained through a data reduction step performed on an initially large matrix. Furthermore this initial data matrix contains far more amino acids than the 20 natural amino acids needed for the creation of PCMs, and therefore the descriptors might not perform optimal within the 20 natural amino acid space needed for description of protein targets. The authors are currently working on a novel protein descriptor that eliminates some of the issues described above.

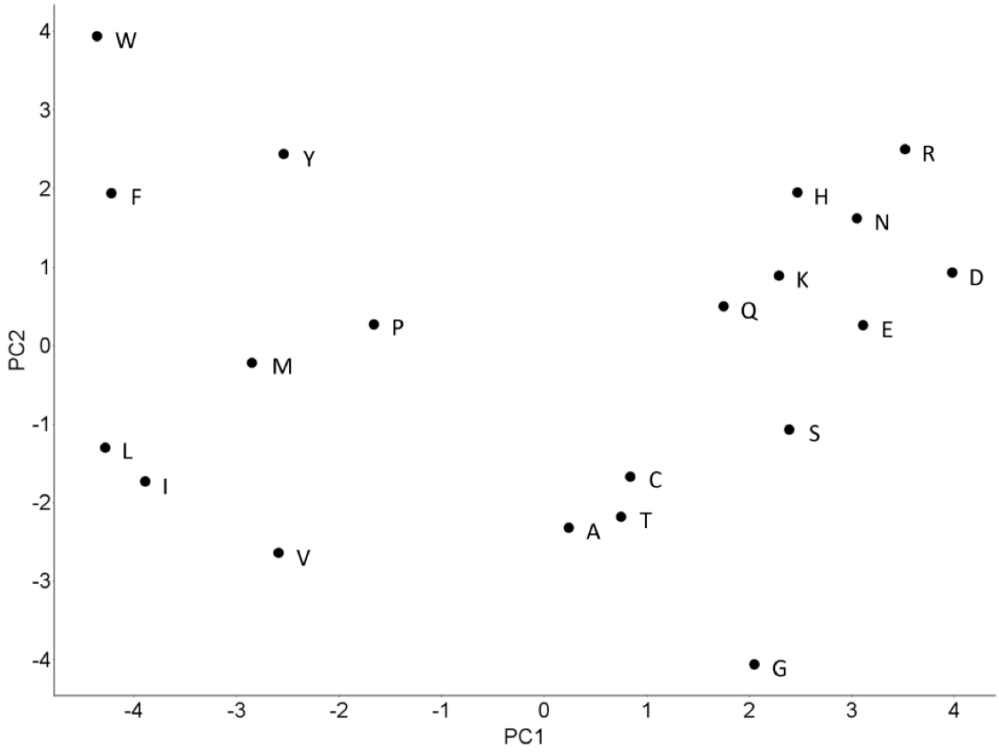


Figure 2.6: Principal components 1 and 2 of the PCA analysis which resulted in the Z-scales. In this plot similar amino acids are located closely together and dissimilar amino acids are spread further apart. PC1 can be interpreted as a lipophilicity scale and PC2 as volume/polarizability scale. While overall distances between amino acids agree with chemical intuition, this is not always the case (such as lysine being located closer to negatively charged amino acids than to arginine, which is reasonable when thinking about locations of the amino acids in the outside of a protein, but less so when relating to ligand binding properties).

2.7 Cross terms

2.7.1 Non-linear term. Descriptor cross terms are used to allow linear machine learning methods like Partial Least Squares (PLS) to model the non-linear interactions that guide ligand-target interactions,[75, 76] which seem to be present in at least half of all structure-activity datasets.[77] Cross terms are influenced by both the ligand and the target part of the dataset and they are intended to model particular interactions between the ligand and the target, such as charge interaction sets.[19, 22] However, when using non-linear machine learning methods in our work in practice, we found that cross terms may in fact deteriorate model performance, likely due to the introduction of a large number of additional parameters, so it is often beneficial to leave them out completely.[57] In their simplest forms, cross terms describe the similarity between different compound – target pairs; therefore they can be constructed as any similarity measure.

2.7.2 Drawbacks. One problem with cross-terms is that their functional form is undefined, and they can be any function of ligand and target properties. In practice often properties of the ligand and target are multiplied.[19, 65] However, this step follows more the intuition of the user than any thorough theoretical derivation. (For a comprehensive overview of previously used cross terms see **Table 2.1**.) Variable selection, mentioned under data pre-processing, can be applied to descriptors prior to cross term calculation when the calculation of all possible cross terms is too time consuming or computationally infeasible.[14, 35]

2.7.3 Alternative approaches. A second and completely different approach to cross term calculation is provided by Lindstrom *et al.*[67] They applied PCM to model a dataset that includes structural protein information, while using the target – ligand interactions as a cross term. The interactions were limited to steric and electrostatic complementarity, the ligand strain energy, logP (octanol/water partition coefficient), steric fit, complementary surface area interactions and the number of rotatable bonds in the ligand as described by Head *et al.*[78] Although their final global model had a prediction error of approximately 1.7 log units, the authors showed that any property dependent on both target and ligand features can be used as a cross term.

Weil and Rognan extend this work by compressing the ligand and target descriptors to a single fixed length bit-string.[29] They described the ligands and targets by a combination of different descriptors and subsequently they merged both the ligand and target bit-strings into one single protein-ligand fingerprint (PLFP). Thereby they used a single descriptor consisting of a ligand and target part to describe the dataset, effectively constructing a model on only the cross term. The authors obtain predictive models with average ROC values of around 0.80 with these descriptors and a variety of machine learning techniques.[29]

In conclusion a cross term can be any descriptor that depends on properties from the compound and properties from the target. As long as this condition is met there is no difference if the cross term is obtained by mathematically combining the compound and target descriptors (e.g. by multiplication, addition or exponentially) or if the cross term is obtained by introduction of an independent descriptor. As far as the authors know, the different mathematical operators to obtain cross terms have not been thoroughly researched, therefore this provides some interesting research opportunities.

2.8 Data pre-processing

Before a viable PCM model can be constructed, the data should usually be pre-processed allowing it to be compatible with the machine learning technique of choice and to obtain optimal model performance. Depending on the data and machine learning technique this processing can be very extensive or relatively simple. Here we will provide an overview of data preprocessing steps, with applications in PCM modeling in mind.

2.8.1 Scaling and mean centering. When a PCM model is created using multiple descriptors (descriptor blocks), it should be prevented that a biased model is created, wherein a subset of descriptors mask the influence of the other descriptors. Scaling and mean centering the different descriptors prevents this bias, makes them compatible and is especially necessary when using PLS modeling.[79] In block scaling one block of descriptors is scaled according to:

$$1/(Nb)1/2 \quad (1)$$

Here, Nb is the number of descriptors in block b. Block scaling prevents larger blocks to mask small ones and it is often incorporated in modern modeling tools.[26, 65] In the case of PCM modeling mean centering is always advised.

2.8.2 Covariance removal. To prevent overfitting of the model, which is likely when PLS is used,[80] it is important to remove covariance within the descriptor blocks before the final model is built. Covariance can lead to misinterpretations of the final model and poor extrapolation capabilities, and it increases the dimensionality of the model with no apparent benefit. Several methods are available for removal of covariance and determination of the descriptors that best describe the data. Two of the covariance removal methods previously used in PCM will be highlighted, namely variable extraction (e.g. PCA) and variable selection. For a general review about the underlying principles and methods see Wegner *et al.*[81]

2.8.3 Variable extraction. An example of variable extraction is PCA, a method to find the underlying latent variables present in a data set, e.g. in a set of descriptors. The original multidimensional space defined by the descriptors is summarized by a smaller number of descriptive dimensions which describe the main variation in the data; these are called the principal components.[82] PLS is the regression extension of PCA and provides the ability to construct a model based on the extracted principal components. PCA can be used when modeling of the original dataset is infeasible due to reasons such as a high dimensionality of the dataset.[14, 65] However, at the same time the interpretability of the model is usually decreased.

2.8.4 Variable selection. Variable selection is the selection and exclusion of variables with negligible importance for the data to be modeled. It is also known as Variable importance projection (VIP) or variable subset selection (VSS). When building a PCM model using PLS, the reliability is heavily influenced by the variable selection before model training. In the case of PLS it can therefore be seen as both a data preprocessing and model tuning procedure.[83] Variable selection is an iterative process aimed at selecting a subset of variables from the full set that optimally describes the variance of the dataset. Two possible forms of variable selection exist.[84] Backward selection consists of iterative elimination starting from the full set, and forward selection consists of iterative addition starting from a single variable. To our knowledge in PCM only backwards selection has been previously used. The iterative process proceeds as follows, firstly a model is constructed on the full data set, subsequently descriptors with negligible importance are removed and a new model on the improved descriptor set is built.[19, 82, 83] This procedure can be repeated until variable selection leads to model deterioration. In models where many variables receive low weights, the variable selection can significantly improve the model.[85]

2.9 Modeling techniques in PCM

Apart from statistical methods, both linear machine learning techniques and non-linear machine learning techniques can be and have been used in PCM. We will describe several approaches, starting with modeling learning techniques already used in PCM and subsequently including new suggestions. We will highlight positive and negative consequences of the different techniques, a short summary of which is given in **Table 2.2**.

Table 2.2: Modeling techniques previously used, and which can potentially be used, in proteochemometric modeling with their main advantages and disadvantages

Technique	Linear ?	Previously used in PCM?	Advantage	Disadvantage
PLS	Linear	Yes	Highly Interpretable	Requires Cross-Terms
RS	Non-linear	Yes	Highly Interpretable	Classification
NN	Non-linear	Yes	Performs well on complex data	High Dimensionality
SVM	Non-linear	Yes	Very Robust on complex data	Poorly Interpretable
NB	Linear	Yes	Performs well on complex data	Requires Cross-Terms
RF	Non-linear	Yes	Low risk of Overfitting	-
DT	Non-linear	Yes	Highly Interpretable	Variable performance
GP	Non-linear	No	Confidence Estimate	Long training time

Explanation of abbreviations: Partial Least Squares (PLS), Rough Set (RS), Neural Net (NN), Support Vector Machines (SVM), Naïve Bayesian (NB), Random Forest (RF), Decision Tree (DT).

2.9.1 Partial least squares. By far the most commonly used modeling technique in PCM is PLS.[80] PLS is the regression extension of PCA and specifies the relationship between an output variable Y and a set of predictor variables Xi. The final model is able to display the role of the individual predictor variables in the model. Advantages of PLS are that it is highly interpretable and requires low computational expense. A large number of PCM models have been created founded on PLS with a prediction error usually ranging between 0.4 and 0.8 log units. Targets for which PLS-PCM models were constructed include melanocortin receptors, HIV Proteases and dengue virus NS3 proteases.[20, 33, 34] However, as PLS is linear it requires cross terms to be calculated in order to obtain an optimal model as opposed to Rough Set modeling which alleviates this restriction.[21, 22, 26, 62, 65, 85, 86]

2.9.2 Rough set modeling. Non-linear Rough Set (RS) modeling has been introduced to PCM by Strömbergsson *et al.* as they proposed that RS might be capable of modeling data to a higher level than PLS because of its non-linearity.[29, 36] RS constitutes a mathematical framework that induces basic IF-THEN decision rules to classify a compound – ligand pair as active or inactive,[27] therefore these models are highly interpretable. However, as RS is a classification tool, RS is unable to perform regression and provide a numerical value, e.g. an affinity (pKi) value. RS modeling has been applied to melanocortin and adrenergic receptor datasets as well as a dataset containing a very broad target collection obtained from the PDB.[27, 35] RS modeling performed very well with an area under the retrieval curve (ROC) usually above 0.9, reliably distinguishing between actives and inactives on a particular target.

However, while the non-linear RS cannot be used in order to model a numerical output variable rather than a class, SVM as described in the following section is a non-linear machine learning technique that is capable of both classification and regression.

2.9.3 Support vector machines. SVM is a non-linear modeling technique also applied multiple times in PCM.[24, 25, 57, 58, 70, 87, 88] The main advantage of this machine learning technique is that it has been proven to be very robust and very capable of modeling QSAR datasets, especially in the case of many dimensions.[89, 90] The disadvantage of SVMs at the moment is the degree to which the models can be interpreted. However, a recent paper by Carlsson *et al.* introduced an approach to improve the interpretation capabilities.[91] When these interpretation methods improve they might lead to SVM models being the machine learning technique of choice in PCM models. Three publications are available that describe PCM modeling based on SVMs. The first one is by Strömbergsson *et al.* who model the entire Enzyme-Ligand interaction space.[70] Although the absolute performance of the PCM model is not very accurate with a prediction error of 1.5 log units, the authors were able to model a very diverse dataset. The second publication by Geppert *et al.* showed recovery rates of active compounds from a database of around 60-70 %,[24] leading to the conclusion that SVM can successfully extrapolate from a combination of ligand and target information to retrieve new active compounds on a related target. This conclusion is supported by Ning *et al.*,[25] who also used multiple assay based models to gain improved performance compared to single assay models.

We share a similar view, having used SVM based PCM models to model an adenosine receptor dataset and an HIV Reverse Transcriptase inhibitor dataset.[58] The SVM models we generated were able to model the data with a prediction error of around 0.5 log units. In conclusion SVM is a robust machine learning technique capable of modeling the complicated PCM data. However, it should be noted that SVM specific parameters like 'gamma' and 'cost' must always be optimized through a proper model selection, for instance by cross validation.[92]

2.9.4 Neural net modeling. The final previously used machine learning technique is neural network (NN) modeling. Fernandez *et al.* have applied NN to PCM modeling using a Bayesian Regularized form of NN.[93] NNs are known for their ability to handle complex input-output relationships and provide robust models of non-linear data. For an extensive review on NN please see Grossberg *et al.*[94] In PCM NNs would not be an optimal machine learning technique since NNs often possess too many parameters for this purpose (as an often mentioned very approximate rule of thumb, when training NNs one should have at least three times more datapoints than variables). In PCM modeling the number of variables is much larger than in QSAR as PCM requires two variables per input dimension. Fernandez *et al.* used a simplified correlation matrix as ultimate input descriptor, keeping the number of input variables low and using Bayesian Regularization to diminish the inherent complexity of the NN. Therefore their model output requires a translation to the original descriptors before the results can be interpreted. Furthermore, NNs are not easily interpretable by themselves. However, Browne *et al.* have described ways of improving the interpretability.[95] A much better interpretable modeling technique known from QSAR models is a Naïve Bayesian classifier. The possibility of using this classifier in PCM models will be described below.

2.9.5 Naïve Bayesian classifier. Naïve Bayesian (NB) classification models have been shown to be able to model datasets with a high number of variables and relate bioactivity predictions from multiple activity classes with each other.[96] These two qualities make Bayesian classification another suitable machine learning technique for PCM. However, as Bayesian classification is linear, cross terms might be required to allow confident modeling. Two publications on PCM with a Naïve Bayesian (NB) classifier have appeared, one by Weil and Rognan and one by Strömbergsson *et al.*[29, 46] However, in the former the model is created merely on cross terms as descriptors and in the second publication no cross terms are used. We speculate that an NB model will reach a better performance and interpretability when separate ligand, protein and cross term descriptors would be used.

2.9.6 Decision trees algorithm. The decision trees (DT) algorithm has been known from the creation of QSAR models. The decision based output makes a decision tree highly interpretable. The DT algorithm has been applied to PCM by Lapins *et al.* and by Strömbergsson *et al.* [39, 46] In the former publication their performance is somewhat disappointing with a squared correlation coefficient of 0.45, which puts it slightly below PLS. The authors contribute this performance to the high non-linearity of the dataset. In the latter publication by Strömbergsson *et al.* the performance of DT, as a classifier, is superior to both SVM and NB based classification models, with an ROC score of 0.84 and an accuracy of 82 %. It can therefore be concluded that the performance of DT can vary with the dataset and that further research is required.

2.9.7 Random forest. Random forest modeling techniques, which can be used for both classification and regression, have previously been shown to perform well on QSAR datasets.[97-99] Weil and Rognan have also applied this technique to PCM data.[29] In their paper RF performs very well on a number of datasets with a recall larger than 0.5 and precision value higher than 0.7 on average, on average RF performs roughly equal to SVM. However, since RF can provide the following additional features: built-in performance assessment, a measure of relative importance of descriptors, and a measure of compound similarity that is weighted by the relative importance of descriptors. RF might very well be a better choice for usage in PCM models than SVM with its low interpretability.

In conclusion, all machine learning methods that have currently been applied to PCM have their advantages and disadvantages, and none can be considered a universal optimal approach. It might therefore be interesting to apply established machine learning methods that have previously not been used in PCM. One suggestions will be discussed below.

2.9.8 Possible new machine learning techniques to be applied in PCM. One of the most promising machine learning techniques not yet applied in PCM are Gaussian Processes (GP).[100] The non-linear GP not only provide the scientist with a prediction of the output variable but also a measure of reliability for this prediction in the form of variance estimation for each prediction. This quality makes it invaluable in PCM as it directly links the application domain to model predictions and allows the selection of the most reliable predictions for decision making. Previously GPs have been shown to perform very well in the modeling of ADME and physicochemical properties.[101, 102] Although the training time required surpasses that of the linear PLS technique, the superior performance combined with the prediction confidence parameter tips the scales in favor of GP.

2.10 Validation of a PCM model

One of the key differences between PCM and QSAR is that PCM significantly increases the number of variables to be modeled as the descriptor space is increased. Therefore there is an increased risk of both chance correlations and model overfitting.[85] The modeler should consequently take care to rule out the possibility of a model built on chance correlations. Validation in PCM is based on established validation techniques normally applied in QSAR modeling. The key goal is to get a reliable estimate of the model quality and applicability. This goal can be achieved by calculation of the correlation coefficient, coefficient of determination and RMSE. For an overview of validation techniques please see a set of recent comprehensive publications.[103-106]

2.10.1 Y-scrambling. Response permutation testing, or Y-scrambling is an approach to estimate the risk of chance correlations.[103, 107] Y-scrambling consists of keeping the X-variables, or the descriptor space, fixed and to randomly shuffle the output or Y- variable and subsequently retraining the model. A typical approach is to create 100 random models and to assess their performance using standard validation parameters and to compare this performance to the actual model, where the performance of the proper model should be significantly higher (for details see above references). Due to the highly increased variable space that a PCM is founded on, this validation technique is very relevant and should always be applied to validate the final model.

2.10.2 Internal validation. Cross validation or internal validation is used to estimate the ability of the model to reliably predict the activity of the data points used in the training set. There are three forms of cross validation, namely Leave-One-Out (LOO)) cross validation, n-fold cross validation and double loop cross validation.[85] Double loop cross validation was introduced by Freyhult *et al.* to improve the quality of cross validation performance estimates and prevent overoptimistic assumptions.

Two independent loops are used to tune model parameters in the inner loop and provide a performance estimate P2 in the outer loop. Here we will only consider n-fold cross validation as it is currently the state of the art.[108] In n-fold cross validation the total training set is divided into 'n' equal subsets. Subsequently a model is trained on n-1 of the subsets and used to predict the activity of the data points in the remaining subset. This process is repeated until all subsets have been left out of the training and the plots of these iterations are used to calculate q^2 and cross-validated RMSE.[109] Cross validation, while useful, cannot always reliably be used to estimate the performance of the final model on unknown data points as it has been shown that there is no direct correlation between q^2 and R^2.[59, 109] However, cross-validation provides a very useful framework for tuning modeling parameters like the number of components in PLS and the values for 'gamma', 'epsilon' and 'cost' in SVM.[65, 89, 92]

2.10.3 External validation. In external validation a trained model is used to predict the output variable for a set of data points for which an observed activity value is available. These points have been separated from the original dataset prior to model training (as well as selection!) or are assembled only after the model has been constructed, and hence they and are completely unknown to the model. Separation of this so called test-set from the training set can be performed using a wide extent of different parameters. (For a full review see Tropsha *et al.*[103]) After model training, the performance of the model on the test set is estimated using conventional validation parameters. The main goal of this exercise is to provide a more reliable performance estimate than internal validation to assess the quality of model predictions on a dataset of unknown compounds. This performance estimation becomes even more critical when using any sort of model selection, e.g. feature selection.[110] Otherwise there is a considerable risk that models will become overfitted as we reported earlier,[111] and is explained by Wegner et al.[81]

2.10.4 Prospective validation. The only true validation for any computational model is prospective validation. Prospective validation assesses model performance by experimentally determining compound activity subsequent to model predictions. Unfortunately in the field of PCM not much work has been published containing a prospective validation. Currently, to the knowledge of the authors, only three publications contain prospective validation of modeling results.

The first prospectively validated PCM model was constructed on a dataset consisting of melanocortin wild-type and chimeric receptors and their ligand α-MSH (and synthetic analogues).[21] In this publication a small scale prospective validation was performed in which PCM predicted a correct response of the binding affinity to point mutations in 80 % of the datapoints. The second published prospective validated PCM model was constructed on a HIV protease dataset (containing point mutations in the protein that change binding affinity) and a selection of octapeptides as ligands.[112] Here PCM was able to prospectively predict the affinity of 10 peptides on 4 different protease mutants with an R of 0.63 (R^2 of 0.40). Lastly, a prospectively validated PCM modeling study was published by Nagamine et al.[113] They used an SVM based model to find androgen receptor ligands from a pool of 19 million compounds while iteratively prospectively validating model predictions. With their model they obtained an area under the ROC curve of 0.717 compared to 0.558 for QSAR. These results firmly establish PCM as a reliable modeling technique with applications in preclinical research.

2.11 Pitfalls and disadvantages

In any form of statistical modeling there are a large number of possible pitfalls which are caused by the modeling technique, data set preparation or can be the result of a simple bias. An extensive review on risks in statistical modeling has been published by Gedeck *et al.*;[13] here we will focus on specific dangers that can arise when performing PCM.

The main risk present in PCM has also been discussed under validation, and is the fact that the large increase in descriptor variables increases the risk of chance correlation models. This should at all times be kept in mind and an extensive validation (including cross-validation, Y-scrambling and prospective testing) is indicated for all PCM models.

A second pitfall comes from the fact that the data to be modeled by PCM is inherently non-linear, whereas many machine learning techniques in QSAR rely on (multiple) linear regression. Therefore the data should be modified in many cases by the introduction of cross-terms. The risk of cross-terms is that they will account for most of the described variance, as cross terms describe variance of the ligands and the targets simultaneously. If this is the case, cross terms can mask the contributions of the pure compound or protein descriptors.

Furthermore, since they rely on both compound and protein contributions, cross terms are not always readily interpretable;[27] however, this depends to a significant extent on the precise way in which cross terms are constructed. When using cross terms one needs to ensure that both compound and protein descriptors are compatible; which is not always the case. A workaround that circumvents the use of cross terms, can be the use of non-linear machine learning techniques; however their low interpretability might lead to highly accurate but non-interpretable black box models. Currently research efforts are underway to alleviate this problem,[91, 95] but for now a decision needs to be made in many cases between better interpretable models or more accurate models. As mentioned above, it also needs to be ensured that cross terms indeed improve model performance, which was not always the case in the experience of the authors.

A disadvantage of PCM is that the calculation time of the models increases compared to QSAR when protein descriptors are added. Depending on the machine learning technique this increase in calculation time can be small, as in PLS, or exponential, as in radial basis function based SVMs.[114]

2.12 Conclusions

PCM is a relatively new technique that, by including target descriptors in addition to ligand descriptors, enables modeling of datasets that could previously only be modeled separately using conventional QSAR based techniques. PCM has been applied successfully to a variety of targets, among which are GPCRs, Viral Proteins and Cytochrome P450 enzymes. However, relatively few of those studies have included prospective validations – with the notable exceptions of the studies by Prusis *et al.*, Kontijevskis *et al.*, and Nagamine *et al*.[21, 112, 113] Hence, while limited data exists, the authors are of the opinion that PCM modeling should indeed, in many cases, be able to make better use of bioactivity data of molecules than previous QSAR models.

In conclusion, by more comprehensively exploiting the information contained in datasets that were previously considered separate, PCM models allow for improved extrapolation both on the ligand side (by taking more 'chemistry' into account) as well as on the target side (by incorporating the relationship between targets into the model). This enables applications in areas such as the deorphanization of compounds or the selection of compounds that are selective, or show a desired bioactivity profile. Until the current stage few prospective PCM studies are available in the literature. However PCM is one of the areas where other research groups are currently active, and where a large-scale validation will be published also by the authors very shortly.

2.13 Acknowledgements

The authors would like to thank Olaf O. van den Hoven for his supporting work and discussions. GJPvW thanks Tibotec BVBA for funding his PhD project.

2.14 References

1. H. Meyer; *Zur theorie der alkoholnarcose.* Arch. Exp. Pathol. Pharmakol.; *1899*. **42**: 109-118.
2. E. Overton; *Studien über die narcose, zugleich ein beitrag zur allgemeinen pharmakologie.* Jena, Gustav Fisher *1901*. **45**: 195.
3. C. Hansch and T. Fujita; *ρ–σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.* J. Am. Chem. Soc.; *1964*. **86** (8): 1616-1626.
4. C. Hansch; *Quantitative approach to biochemical structure-activity relationships.* Acc. Chem. Res.; *1969*. **2** (8): 232-239.
5. D.E. Clark; *What has computer-aided molecular design ever done for drug discovery?* Expert Opin. Drug Discov.; *2006*. **1** (2): 103-110.
6. J.A. DiMasi, R.W. Hansen, and H.G. Grabowski; *The price of innovation: new estimates of drug development costs.* Journal of Health Economics; *2003*. **22** (2): 151-185.
7. A. Bender and R.C. Glen; *A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication.* J. Chem. Inf. Model.; *2005*. **45** (5): 1369-1375.
8. A. Bender and R.C. Glen; *Molecular similarity: a key technique in molecular informatics.* Org. Biomol. Chem.; *2004*. **2**: 3204-3218.
9. T. Klabunde; *Chemogenomic approaches to drug discovery: similar receptors bind similar ligands.* Br. J. Pharmacol.; *2007*. **152** (1): 5-7.
10. D. Rognan; *Chemogenomic approaches to rational drug design.* Br. J. Pharmacol.; *2007*. **152**: 38-52.
11. M. Lapinsh, P. Prusis, et al.; *QSAR and Proteo-chemometric Analysis of the Interaction of a Series of Organic Compounds with Melanocortin Receptor Subtypes.* J. Med. Chem.; *2003*. **46** (13): 2572-2579.
12. L.M. Kauvar; *Affinity fingerprinting.* Biotechnology (N Y); *1995*. **13** (9): 965-966.
13. P. Gedeck, B. Rohde, and C. Bartels; *QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets.* J. Chem. Inf. Model.; *2006*. **46** (5): 1924-1936.

14. M. Lapinsh, P. Prusis, et al.; *Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions.* Bioinformatics; *2005*. **21** (23): 4289-4296.

15. A.F. Fliri, W.T. Loging, et al.; *Biospectra Analysis: Model Proteome Characterizations for Linking Molecular Structure and Biological Response.* J. Med. Chem.; *2005*. **48** (22): 6918-6925.

16. R. Guha and J.H. VanDrie; *Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs.* J. Chem. Inf. Model.; *2008*. **48** (3): 646-658.

17. J.L. Medina-Franco, K. Martinez-Mayorga, et al.; *Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs.* J. Chem. Inf. Model.; *2009*. **49** (2): 477-491.

18. M. Wawer, L. Peltason, and J.r. Bajorath; *Elucidation of Structureâˆ'Activity Relationship Pathways in Biological Screening Data.* J. Med. Chem.; *2009*. **52** (4): 1075-1080.

19. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.

20. M. Lapinsh, S. Veiksina, et al.; *Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes.* Mol. Pharmacol.; *2005*. **67** (1): 50 - 59.

21. P. Prusis, S. Uhlén, et al.; *Prediction of indirect interactions in proteins.* BMC Bioinformatics; *2006*. **7**: 167-180.

22. J.E.S. Wikberg, F. Mutulis, et al.; *Melanocortin receptors: Ligands and proteochemometrics modeling;* in *Melanocortin System*; D. Braaten; Editor *2003*: New York. p. 21-26.

23. E. Van der Horst, J.E. Peironcely, et al.; *Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1964-1977.

24. H. Geppert, J. Humrich, et al.; *Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors.* J. Chem. Inf. Model.; *2009*. **49** (4): 767-779.

25. X. Ning, H. Rangwala, and G. Karypis; *Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets.* J. Chem. Inf. Model.; *2009*. **49** (11): 2444-2456.

26. M. Lapinsh, P. Prusis, et al.; *Proteochemometric modeling reveals the interaction site for Trp9 modified alpha-MSH peptides in melanocortin receptors.* Proteins: Struct., Funct., Bioinf.; *2007*. **67** (3): 653-660.

27. H. Strombergsson, P. Prusis, et al.; *Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions.* Proteins: Struct., Funct., Bioinf.; *2006*. **63** (1): 24-34.

28. P. Prusis, R. Muceniece, et al.; *PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions.* Biochim. Biophys. Acta; *2001*. **1544**: 350 - 357.

29. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.

30. J.R. Bock and D.A. Gough; *Virtual screen for ligands of orphan G protein-coupled receptors.* J. Chem. Inf. Model.; *2005*. **45** (5): 1402-1414.

31. L. Jacob, B. Hoffmann, et al.; *Virtual screening of GPCRs: An in silico chemogenomics approach.* BMC Bioinformatics; *2008*. **9** (1): 363-379.

32. M. Lapins, M. Eklund, et al.; *Proteochemometric modeling of HIV protease susceptibility.* BMC Bioinformatics; *2008*. **9** (1): 181-192.

33. M. Lapins and J.E.S. Wikberg; *Proteochemometric Modeling of Drug Resistance over the Mutational Space for Multiple HIV Protease Variants and Multiple Protease Inhibitors.* J. Chem. Inf. Model.; *2009*. **49** (5): 1202-1210.

34. P. Prusis, M. Lapins, et al.; *Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases.* Bioorg. Med. Chem.; *2008*. **16** (20): 9369-9377.

35. H. Strombergsson, A. Kryshtafovych, et al.; *Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures.* Proteins: Struct., Funct., Bioinf.; *2006*. **65** (3): 568-579.

36. I. Mandrika, P. Prusis, et al.; *Proteochemometric modelling of antibody-antigen interactions using SPOT synthesised peptide arrays.* Protein Eng., Des. Sel.; *2007*. **20** (6): 301 - 307.

37. B. Pirard and H. Matter; *Matrix metalloproteinase target family landscape: a chemometrical approach to ligand selectivity based on protein binding site analysis.* J. Med. Chem.; *2006*. **49** (1): 51-69.

38. M. Fernandez, S. Ahmad, and A. Sarai; *Proteochemometric Recognition of Stable Kinase Inhibition Complexes Using Topological Autocorrelation and Support Vector Machines.* J. Chem. Inf. Model.; *2010*. **50** (6): 1179-1188.

39. M. Lapins and J. Wikberg; *Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques.* Bmc Bioinformatics; *2010*. **11** (1): 339.

40. I. Dimitrov, P. Garnev, et al.; *Peptide binding to the HLA-DRB1 supertype: A proteochemometrics analysis.* Eur. J. Med. Chem.; *2010*. **45** (1): 236-243.

41. A. Kontijevskis, J. Komorowski, and J.E.S. Wikberg; *Generalized Proteochemometric Model of Multiple Cytochrome P450 Enzymes and Their Inhibitors.* J. Chem. Inf. Model.; *2008*. **48** (9): 1840-1850.

42. J.R. Bock and D.A. Gough; *A New Method to Estimate Ligand-Receptor Energetics.* Molecular & Cellular Proteomics; *2002*. **1** (11): 904-910.

43. A. Christopoulos; *Allosteric binding sites on cell-surface receptors: novel targets for drug discovery.* Nat. Rev. Drug Discovery; *2002*. **1** (3): 198-210.

44. Z.G. Gao; *Allosteric modulation of the adenosine family of receptors.* Mini reviews in medicinal chemistry; *2005*. **5** (6): 545.

45. V.J. Merluzzi, K.D. Hargrave, et al.; *Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor.* Science; *1990*. **250** (4986): 1411-1413.

46. H. Strömbergsson, M. Lapins, et al.; *Towards Proteome–Wide Interaction Models Using the Proteochemometrics Approach.* Molecular Informatics; *2010*. **29** (6-7): 499-508.

47. W. Soudijn, I. van Wijngaarden, and A.P. IJzerman; *Allosteric modulation of G protein-coupled receptors: perspectives and recent developments.* Drug Discov. Today; *2004*. **9** (17): 752-758.

48. A. Bender, J.L. Jenkins, et al.; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.

49. R. Todeschini and V. Consonni; *Handbook of Molecular Descriptors 2000*; Weinheim: WILEY-VCH.

50. N. Wale, I. Watson, and G. Karypis; *Comparison of descriptor spaces for chemical compound retrieval and classification.* Knowledge and Information Systems; *2008*. **14** (3): 347-375.

51. J. MacCuish, C. Nicolaou, and N.E. MacCuish; *Ties in Proximity and Clustering Compounds.* J. Chem. Inf. Comput. Sci.; *2000*. **41** (1): 134-146.

52. D.M. Hawkins; *The Problem of Overfitting.* J. Chem. Inf. Comput. Sci.; *2003*. **44** (1): 1-12.

53. E. Van der Horst and A.P. IJzerman; *Computational Approaches to Fragment and Substructure Discovery and Evaluation;* in *Fragment-Based Drug Discovery: A Practical Approach*; E.R. Zartler and M.J. Shapiro; Editors. *2008*; John Wiley & Sons, Ltd: Chichester, West Sussex, U.K.

54. R.C. Glen, A. Bender, et al.; *Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME.* IDrugs; *2006*. **9** (3): 199 - 204.

55. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.

56. A. Bender, H.Y. Mussa, et al.; *Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance.* J. Chem. Inf. Comput. Sci.; *2004*. **44** (5): 1708-1718.

57. M.R. Doddareddy, G.J.P. van Westen, et al.; *Chemogenomics: Looking at biology through the lens of chemistry.* Statistical Analysis and Data Mining; *2009*. **2** (3): 149-160.

58. G.J.P. Van Westen, J.K. Wegner, et al.; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

59. H. Kubinyi, F.A. Hamprecht, and T. Mietzner; *Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices*. J. Med. Chem.; *1998*. **41** (14): 2553-2564.

60. T. Scior, J.L. Medina-Franco, et al.; *How to Recognize and Workaround Pitfalls in QSAR Studies:A Critical Review.* Curr. Med. Chem.; *2009*. **16**: 4297-4313.

61. M. Pastor, G. Cruciani, et al.; *GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors.* J Med Chem; *2000*. **43**: 3233 - 3243.

62. M. Lapinsh, P. Prusis, et al.; *Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.* Mol. Pharmacol.; *2002*. **61** (6): 1465-1475.

63. K. Ye, E.W.M. Lameijer, et al.; *A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **63** (4): 1018-1030.

64. S. Hellberg, M. Sjoestroem, et al.; *Peptide quantitative structure-activity relationships, a multivariate approach.* J. Med. Chem.; *1987*. **30** (7): 1126-1135.

65. A. Kontijevskis, R. Petrovska, et al.; *Proteochemometric analysis of small cyclic peptides' interaction with wild-type and chimeric melanocortin receptors.* Proteins: Struct., Funct., Bioinf.; *2007*. **69** (1): 83-96.

66. G.J.P. van Westen, J.K. Wegner, et al.; *Mining protein dynamics from sets of crystal structures using "consensus structures".* Protein Sci.; *2010*. **19** (4): 742-752.

67. A. Lindström, F. Pettersson, et al.; *Hierarchical PLS Modeling for Predicting the Binding of a Comprehensive Set of Structurally Diverse Protein−Ligand Complexes.* J. Chem. Inf. Model.; *2006*. **46** (3): 1154-1167.

68. T.R. Hvidsten; *A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins.* Bioinformatics; *2003*. **19** (2): 81.

69. T.R. Hvidsten, A. Kryshtafovych, and K. Fidelis; *Local descriptors of protein structure: A systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions.* Proteins: Struct., Funct., Bioinf.; *2009*. **75** (4): 870-884.

70. H. Strombergsson, P. Daniluk, et al.; *Interaction Model Based on Local Protein Substructures Generalizes to the Entire Structural Enzyme-Ligand Space.* J. Chem. Inf. Model.; *2008*. **48** (11): 2278-2288.

71. A.G. Georgiev; *Interpretable numerical descriptors of amino acid space.* J. Comput. Biol.; *2009*. **16** (5): 703-723.

72. H. Mei, Z.H. Liao, et al.; *A new set of amino acid descriptors and its application in peptide QSARs.* Biopolymers; *2005*. **80** (6): 775-786.

73. M. Sandberg, L. Eriksson, et al.; *New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids.* J. Med. Chem.; *1998*. **41** (14): 2481-2491.

74. A. Zaliani and E. Gancia; *MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies.* J. Chem. Inf. Comput. Sci.; *1999*. **39** (3): 525-533.

75. D.H. Williams, N.L. Davies, et al.; *Noncovalent Interactions: Defining Cooperativity. Ligand Binding Aided by Reduced Dynamic Behavior of Receptors. Binding of Bacterial Cell Wall Analogues to Ristocetin A.* J. Am. Chem. Soc.; *2004*. **126** (7): 2042-2049.

76. A.D. Williams, S. Shivaprasad, and R. Wetzel; *Alanine Scanning Mutagenesis of A[beta](1-40) Amyloid Fibril Stability.* Journal of molecular biology; *2006*. **357** (4): 1283-1294.

77. Y. Patel, V.J. Gillet, et al.; *Assessment of Additive/Nonadditive Effects in Structure-Activity Relationships: Implications for Iterative Drug Design.* J. Med. Chem.; *2008*. **51** (23): 7552-7562.

78. R.D. Head, M.L. Smythe, et al.; *VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands.* J. Am. Chem. Soc.; *1996*. **118** (16): 3959-3969.

79. S. Wold, M. Sjöström, and L. Eriksson; *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems; *2001*. **58** (2): 109-130.

80. P. Geladi and B. Kowalski; *Partial least-squares regression: a tutorial.* Anal. Chim. Acta; *1986*. **185**: 1.

81. J.K. Wegner, H. Froehlich, and A. Zell; *Feature Selection for Descriptor Based Classification Models. 1. Theory and GA-SEC Algorithm.* J. Chem. Inf. Comput. Sci.; *2004*. **44** (3): 921-930.

82. L. Eriksson, P.L. Andersson, et al.; *Megavariate analysis of environmental QSAR data. Part I--a basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD).* Mol. Diversity; *2006*. **10**: 169-186.

83. A. Hoskuldsson; *Variable and subset selection in PLS regression.* Chemometrics and Intelligent Laboratory Systems; *2001*. **55** (1-2): 23-38.

84. I. Guyon and A. Elisseeff; *An introduction to variable and feature selection.* J. Mach. Learn. Res.; *2003*. **3**: 1157-1182.

85. E. Freyhult, P. Prusis, et al.; *Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling.* BMC Bioinformatics; *2005*. **6** (50).

86. H. Sun and D. Fry; *Molecular Modeling of Melanocortin Receptors.* Curr. Top. Med. Chem.; *2007*. **7**: 1042-1051.

87. C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*. 2001; Available from: http://www.csie.ntu.edu.tw/cjlin/libsvm.

88. C. Cortes and V. Vapnik; *Support-vector networks.* Machine Learning; *1995*. **20** (3): 273-297.

89. X.J. Yao, A. Panaye, et al.; *Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression.* J. Chem. Inf. Comput. Sci.; *2004*. **44** (4): 1257-1266.

90. Y. Liu; *A comparative study on feature selection methods for drug discovery.* J. Chem. Inf. Comput. Sci.; *2004*. **44** (5): 1823-1828.

91. L. Carlsson, E.A. Helgee, and S. Boyer; *Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data.* J. Chem. Inf. Model.; *2009*. **49** (11): 2551-2558.

92. A.J. Smola and B. Schölkopf; *A tutorial on support vector regression.* Statistics and Computing; *2004*. **14** (3): 199-222.

93. M. Fernandez, L. Fernandez, et al.; *Proteochemometric Modeling of the Inhibition Complexes of Matrix Metalloproteinases with N-Hydroxy-2-[(Phenylsulfonyl)Amino]Acetamide Derivatives Using Topological Autocorrelation Interaction Matrix and Model Ensemble Averaging.* Chem. Biol. Drug Des.; *2008*. **72** (1): 65-78.

94. S. Grossberg; *Nonlinear neural networks: Principles, mechanisms, and architectures.* Neural Networks; *1988*. **1** (1): 17-61.

95. A. Browne, B.D. Hudson, et al.; *Biological data mining with neural networks: implementation and application of a flexible decision tree extraction algorithm to genomic problem domains.* Neurocomputing; *2004*. **57**: 275-293.

96. A. Bender, D.W. Young, et al.; *Chemogenomic Data Analysis: Prediction of Small-Molecule Targets and the Advent of Biological Fingerprints.* Combinatorial Chemistry & High Throughput Screening; *2007*. **10**: 719-731.

97. L. Breiman; *Random Forests.* Machine Learning; *2001*. **45** (1): 5-32.

98. M.R. Segal *Machine Learning Benchmarks and Random Forest Regression*. UC San Francisco: Center for Bioinformatics and Molecular Biostatistics.; *2004*.

99. V. Svetnik, A. Liaw, et al.; *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.* J. Chem. Inf. Comput. Sci.; *2003*. **43** (6): 1947-1958.

100. C.E. Rasmussen; *Gaussian Processes in Machine Learning;* in *Advanced Lectures on Machine Learning 2004*. p. 63-71.

101. O. Obrezanova, G. Csanyi, et al.; *Gaussian Processes:A Method for Automatic QSAR Modeling of ADME Properties.* J. Chem. Inf. Model.; *2007*. **47** (5): 1847-1857.

102. T. Schroeter, A. Schwaighofer, et al.; *Predicting Lipophilicity of Drug-Discovery Molecules using Gaussian Process Models.* ChemMedChem; *2007*. **2** (9): 1265-1267.

103. A. Tropsha, P. Gramatica, and Vijay K. Gombar; *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.* QSAR Comb. Sci.; *2003*. **22** (1): 69-77.

104. L. Eriksson; *Quantitative structure-activity relationship validation.* Quantitative structure-activity relationships in environmental sciences-VII SETAC, Pensacola; *1997*: 381 - 397.

105. L. Eriksson and E. Johansson; *Multivariate design and modeling in QSAR.* Chemometrics and Intelligent Laboratory Systems; *1996*. **34** (1): 1.

106. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

107. L. Eriksson, J. Jaworska, et al.; *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environ. Health Perspect.; *2003*. **111** (10): 1361-1375.

108. K. Baumann; *Cross-validation as the objective function for variable-selection techniques.* TrAC Trends in Analytical Chemistry; *2003*. **22** (6): 395-406.

109. A. Golbraikh and A. Tropsha; *Beware of q2!* Journal of Molecular Graphics and Modelling; *2002*. **20** (4): 269-276.

110. J. Reunanen; *Overfitting in making comparisons between variable selection methods.* J. Mach. Learn. Res.; *2003*. **3**: 1371-1382.

111. J.K. Wegner and A. Zell; *Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method.* J. Chem. Inf. Comput. Sci.; *2003*. **43** (3): 1077-1084.

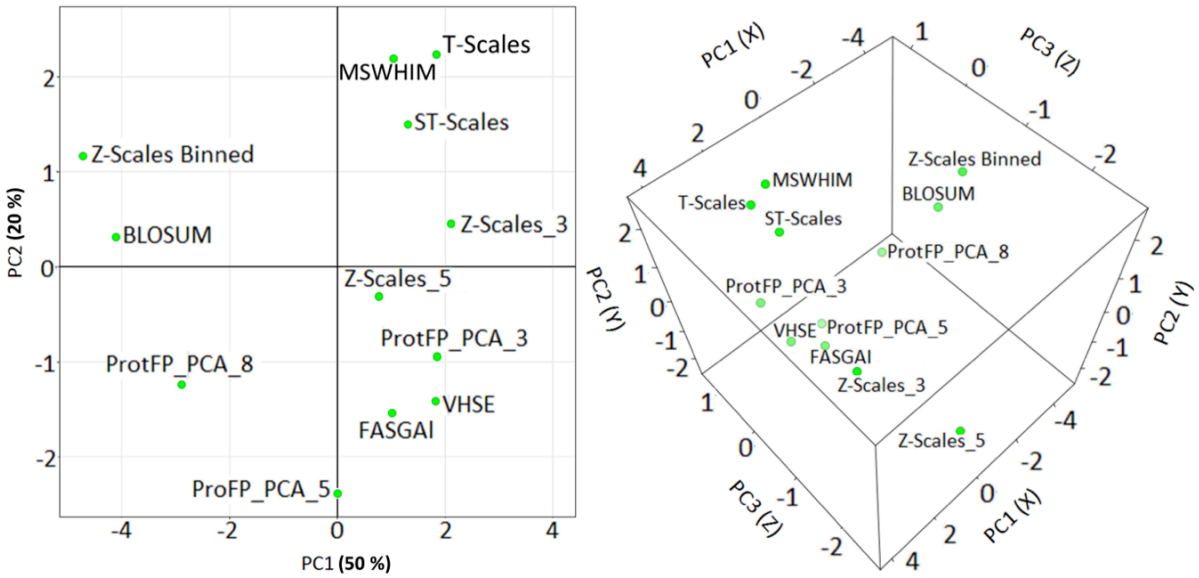
112. A. Kontijevskis, R. Petrovska, et al.; *Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates.* Bioorg. Med. Chem.; *2009*. **17** (14): 5229-5237.

113. N. Nagamine, T. Shirakawa, et al.; *Integrating Statistical Predictions and Experimental Verifications for Enhancing Protein-Chemical Interaction Predictions in Virtual Screening.* PLoS Comput. Biol.; *2009*. **5** (6): e1000397.

114. B. Schölkopf; *Learning With Kernels2002*.

Chapter 3

Comparative Study and Benchmarking of 13 Amino Acids Descriptors and Applications to Proteochemometric Modeling



G.J.P. Van Westen, R.F. Swier, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender.

(Manuscript submitted)

Contents

3.1 Abstract

While a large body of work exists on comparing and benchmarking descriptors of molecular structures, a similar comparison on protein descriptors has not yet been performed. Hence, in the current work a total of 13 different protein descriptor sets have been compared with respect to their behavior in perceiving similarities between amino acids, and benchmarked with respect to their ability of establishing bioactivity models. We investigate which descriptors show complementarities in behavior via principal component analysis, and secondly evaluate prediction performance in five structure-activity benchmarks. These comprise one Angiotensin Converting Enzyme inhibitor data (dipeptides), and two proteochemometric data sets (GPCR ligands and multiple GPCRs; enzyme inhibitors and multiple mutants). In describing amino acid similarities, MSWHIM, T-scales and ST-scales show similar behavior, as do VHSE, FASGAI, and ProtFP_PCA (3). The ProtFP_PCA (5), ProtFP_PCA (8), Z-Scales (Binned), and BLOSUM descriptor sets show behavior that is distinct from another and the clusters above. The use of more principal components (>3 per amino acid) leads to a significant difference in the way amino acids are described, despite capturing less variation of the original input data. In bioactivity modeling protein descriptors perform similar (< 0.2 log units RMSE difference), while the performance per protein is still highly variable. T-scales perform the best overall, while one of our ProtFP descriptors performed the worst. Here we provide a comparison of how similar (and different) currently available descriptor sets perceive amino acids to be. We conclude that in a given situation amino acid descriptors from the different clusters should be explored. This is consistent with our observation that while the performance of modeling bioactivity data using different descriptors is overall relatively similar, some descriptors still perform much better than other descriptors on a particular dataset.

3.2 Introduction

3.2.1 Proteochemometric modeling. Proteochemometric (PCM) modeling uses statistical modeling techniques to model the ligand – target space.[1-4] Related to Quantitative Structure-Activity Relationship (QSAR) modeling, PCM modeling takes both ligand- and target space into account, enabling the models to extrapolate (within limits imposed by the data sets, the descriptors and the modeling method) in both the chemical (ligand) as well as the biological (target) domain. Possible applications include receptor deorphanization, virtual screening for compounds that are selective for a single member of a target family (e.g. the adenosine receptor family), and combined modeling of orthosteric and allosteric compounds (e.g. nucleoside and non-nucleoside HIV reverse transcriptase inhibitors).[3] Hence, the target description is as important as the ligand description. While several publications are available using varying ligand descriptors, on the side of target description there is less literature available, a void we wanted to fill with the current work.[5-7] Previous PCM modeling has been performed using peptide descriptors obtained from the field of Quantitative Sequence-Activity Modeling (QSAM),[2, 8, 9] but later techniques also used different approaches in target description which did not rely on the target sequence (as is the case with QSAM descriptors) but are more structural (e.g. oriented more towards spatial descriptions of the binding site or based on known ligand – target interactions).[10-12]

3.2.2 Utilization of Quantitative Sequence Activity Modeling (QSAM) derived descriptor sets. QSAM attempts to quantitatively explain binding affinity of small peptide drugs to protein (or, more generally, macromolecular) targets, similar to QSAR in the field of small molecules and in this context several descriptor sets for amino acids (AAs) have been developed.[13] The majority of these descriptor sets rely on a principal component analysis (PCA) of a large property matrix used to describe the individual AAs. The data is then reduced in dimensionality *via* PCA while still describing typically over 80% of the variation present in the original set.[9] In general this leads to descriptor sets that can correlate peptide make-up with an output variable (as long as this output variable can be described in terms of individual AA properties in the first place). However, Z-scales, the most widely used descriptor set in PCM modeling was intended to be used in research for small peptide drugs and, hence, covers also non-natural AAs. This is also true for the T-scales and ST-scales. Therefore, if the original matrix consists of over 167 AAs (ST-scales) out of which only 20 are natural AAs, then it is not directly clear how large the fraction of the 'AA property space' is formed by the natural amino acids in respect to the total property space.

Hence this leads to potentially less resolution in the space we are particularly interested in modeling accurately, namely the space formed by the natural amino acids.[14] This balance between non-natural and natural AAs leads to principal components after data reduction that are not necessarily the most relevant ones to describe the natural AAs and, hence, previously developed peptide descriptor sets might not be ideal for use in PCM models. In order to capture the current state-of-the-art in describing AA (and peptide) properties, and to potentially improve upon the current situation, in this work we have benchmarked 13 previously published and four novel AA descriptor sets in order to evaluate the performance of QSAM descriptor sets in PCM.

3.2.3 Amino acid descriptor sets considered in this study. In the current work we have benchmarked a total of 13 different individual descriptor sets where the AA descriptor sets used belong to different broad classes (**Table 3.1**). Firstly, three descriptor sets, namely Z-scales (all versions), VHSE and ProtFP PCA (all versions), are based on a PCA analysis of physicochemical properties.[9, 15] Secondly, ST-scales and T-scales consist of a principal component analysis of mostly topological properties.[14, 16] FASGAI, part of the third category of descriptor tested is based on a factor analysis of physicochemical properties.[17] Furthermore, we also tested two descriptor sets that are calculated in a very different manner compared to the first six, namely a descriptor set based on three dimensional electrostatic properties calculated per AA (MS-WHIM).[18] Additionally, a descriptor set based on a VARIMAX analysis of physicochemical properties which were subsequently converted to indices based on the BLOSUM62 substitution matrix (BLOSUM).[19] Finally, we tested a descriptor set only describing each AA by a single feature (ProtFP Feature).[20, 21] See **Table 3.1** for an overview.

Table 3.1. Descriptor sets included.

Descriptor Set	Type	Derived by	# of components	Variance explained	AAs Covered
BLOSUM	Physicochemical and substitution matrix	VARIMAX	10	n/a	20
FASGAI	Physicochemical	Factor Analysis	6	84%	20
MSWHIM	3D electrostatic potential	PCA	3	61%	20
ProtFP (3)	Physicochemical	PCA	3	75%	20
ProtFP (5)	Physicochemical	PCA	5	83%	20
ProtFP (8)	Physicochemical	PCA	8	92%	20
ProtFP (Feature)	Feature based	Hashing	n/a	n/a	20
ST-scales	Topological	PCA	5	91%	167
T-scales	Topological	PCA	8	72%	135
VHSE	Physicochemical	PCA	8	77%	20
Z-scales (3)	Physicochemical	PCA	3	n/a	87
Z-scales (5)	Physicochemical	PCA	5	87%	87
Z-scales (Binned)	Physicochemical	PCA followed by binning	n/a	n/a	20

The first column contains the name of the descriptor set as used in the main text. The last column differentiates between descriptor sets only covering the natural amino acids or more. Not available is abbreviated by n/a.

3.2.4 Summary of the comparative study of AA descriptor sets and benchmarking. In the current work we characterize the similarity of amino acids, as perceived by each descriptor set considered in this study. Furthermore we benchmark all descriptor sets on three different data sets by constructing structure-bioactivity models and comparing their performance. The datasets are firstly a previously published set of 58 dipeptides that have an inhibitory effect on the angiotensin-converting enzyme (ACE);[22] secondly, a set of 26 GPCRs and approximately 100 active and 100 inactive compounds per receptor obtained from ChEMBL 11;[23] and finally, a set of 451 non-nucleoside reverse transcriptase inhibitors (NNRTIs) and 14 HIV mutants which was also used in a previous publication (but where the protein descriptor used was not varied).[21] It is our hypothesis that the descriptor sets based on solely the natural AAs outperform the descriptor sets created for QSAM work.

3.3 Materials and Methods

A more detailed outline of each descriptor set, illustrating the differences and similarities between all of them, is given below. For each descriptor set a short name used in the tables and figures is given in parentheses.

3.3.1 Z-scales. Z-scales are based on physicochemical properties of the AAs including NMR data and thin-layer chromatography data. Sandberg *et al.*[9] improved on the original Z-scales published by Hellberg *et al.*[24] by introducing two more Z-scales bringing the total to five scales rather than three. Sandberg *et al.* used 26 properties from 87 AAs. The PCA mainly captures lipophilicity (Z1), bulk (Z2), electrogenicity (Z3). The fourth and fifth scale (Z4 and Z5) are more difficult to interpret relating to properties as electronegativity, heat of formation, electrophilicity and hardness. The total variance explained by these five components is 87 %.

In this study we employed the Z-scales using 5 scales (***Z-scales (5)***) and the Z-scales using 3 scales (***Z-scales (3)***) both of which have been used in previous work.[25] Furthermore, the 5 Z-scales were also binned into several classes per scale (***Z-scales (Binned)***). When an AA fell within one of these bins, the bin property was set '1', otherwise it was set '0'. All natural amino acids were uniquely identifiable based on the classification.

For instance Tryptophan is assigned a '1'for the following classes: Lipophilicity High, Size Large, Electronic Properties High, Electronegativity High and Electrophilicity Low, whereas Glycine is assigned a '1' for the following: Lipophilicity Low, Size Small, Electronic Properties High, Electronegativity Medium Low and Electrophilicity Medium Low. The rationale was that these descriptors would be easier to interpret than descriptors derived from a PCA (see Supporting **Table S1** for the classes).

3.3.2 Vectors of Hydrophobic, Steric, and Electronic properties (VHSE). Originally published by Mei *et al.***,** Vectors of Hydrophobic, Steric, and Electronic properties (***VHSE***) are obtained from 18 hydrophobic, 17 steric and 15 electronic properties, giving rise to a total of 50 physicochemical properties of the 20 natural AAs.[15] For each of these three categories a PCA was generated and resulted in Principal Components (PC) of two hydrophobic, two steric and four electronic properties with a total variance of 74.33%, 78.68% and 77.97%, respectively. These eight properties form the VHSE scales.[15]

3.3.3 T-scales. Published by Tian *et al.*, the T-scale descriptor (***T-scales***) is derived from several computer programs utilized to generate 67 common topological descriptors of 135 AAs.[16] These topological descriptors are one of the most simplified descriptors since they are derived from an atom-connecting manner in 2D structures of molecules and therefore do not need an optimization of the 3D structures. A PCA calculation of the five most representative descriptors was called the T scales. These five descriptors encompass 91.14% of the total variance of the data.[16]

3.3.4 ST-scales. Published by Yang *et al.*, The topological ST-scale (***ST-scales***) descriptor is very similar to the T-scales, extending it by taking 827 properties into account which are mainly constitutional, topological, geometrical, hydrophobic, electronic, and steric properties of a total of 167 AAs.[14] For the ST-scales the molecular structures were first optimized as some of the properties used are conformation-dependent. ST-scale utilizes eight PCs instead of the five PCs of T-scales and describes 71.5% of the total variance of the data.[14]

3.3.5 MS-WHIM. Previously published by Zaliani and Gancia, the MS-Whim (***MSWHIM***) descriptor set is derived from 36 electrostatic potential properties derived from the three-dimensional structure of the molecule.[18] These are calculated from 12 statistical parameters starting from x, y, z coordinates of the Connolly surface, which is a solvent-excluded surface (an inverse solvent-accessible surface).[26] On these 36 parameters (3 coordinates by 12 parameters each) of the 20 natural AAs a PCA was performed which gave rise to a set of 3 principal components with a total variance of 61%, as well as a set of 7 principal components with a total of variance of 87%.

However according to the loading plots, the authors concluded that the most representative values were contained in the first three principal components and they hence chose to take only the first three principal components into account in their final descriptor set.[18]

3.3.6 Factor Analysis Scales of Generalized Amino Acid Information (FASGAI). Published by Guizhao and Zhiliang**,** Factor Analysis Scales of Generalized AA Information (***FASGAI***) is derived from 335 physicochemical properties of the 20 natural AAs.[17] Contrary to the other descriptor sets a factor analysis is applied rather than a PCA. Factor analysis also simplifies large quantities of data like PCA does, however factor analysis computes a smaller number of factors that describe the *correlated* variables, whereas PCA searches for the parameters with the largest *variance*. After generating these factors, a PCA was applied to get the factors that would describe the data with the most variance. The PCA resulted in the FASGAI protein descriptor of 6 principal components with a total variance of 83.5%.[17]

3.3.7 BLOSUM. Published by Georgiev, the BLOSUM matrix-derived amino acid descriptors (***BLOSUM***) is the only AA descriptor set we employed that is not directly based on physical or chemical properties of the AAs, but on both physicochemical properties that have been subjected to a VARIMAX analyses and an alignment matrix of the 20 natural AAs, the BLOSUM62 matrix (for details see the work by Georgiev).[19, 27] This procedure renders scales analogous to the Z-scales.[19] This descriptor was added due to its fundamentally different nature and an anticipated complementarity in capturing AA properties, compared to other descriptor sets.

3.3.8 Protein Fingerprint (ProtFP). In addition to the previously published descriptor sets, we also employed a novel AA descriptor set in this work which we termed 'Protein Fingerprint' ('ProtFP'). ProtFP is based on a selection of different AA properties obtained from the AAindex database.[28] However, the difference to descriptor sets mentioned previously is that we started with the full set of indices, while repetitively removing indices with the highest covariance. The final descriptor comes in several flavors. The first ProtFP descriptor (described in more detail below) is based on a PCA of the remaining indices employing 3, 5 or 8 principal components (***ProtFP_PCA (3)***, ***ProtFP_PCA (5)*** or ***ProtFP_PCA (8)***), which allows for quantitative comparison of AAs.

The second variation is based on a hashing approach of all indices values per AA (***ProtFP Feature***), as we published previously.[20, 21] Given the novelty of the ProtFP descriptor sets, their derivation is described in more detail in the following.

3.3.9 Selection of AAindices (for ProtFP). The ProtFP descriptor set was constructed from a large initial selection of indices obtained from the AAindex database for all 20 naturally occurring AAs. This is a principal difference to several other AA descriptor sets, where also non-natural AAs were taken into account.[28] Covariance between indices was determined *via* PCA and indices were normalized and scaled to a range between 0 and 1 rather than using the raw indices. The analysis was performed using the Pipeline Pilot implementation, version 6.1.5, of R-statistics and the 'prcomp' package, with the options of 'mean centering' and 'scaling' enabled.[29] Indices showing highest covariance were removed, while at the same time a number of largely independent physicochemical parameters were maintained. The final reduced selection consisted of 58 AAindices, which are hence (a) based on the relevant natural amino acids only, (b) largely independent (since those indices with larges covariance were removed). The final amino acid indices employed in the construction of the ProtFP descriptor set are listed in Supporting **Table S2**.

3.3.10 PCA of final indices selection (for ProtFP_PCA). In order to obtain descriptors at lower dimensionality PCA was performed on the final set of 58 amino acid properties. The analysis was performed using default parameters, requiring a minimum explained variance of 75%, but forcing a minimum of 8 principal components (PCs). The first three PCs explained 75% of the variance, 5 PCs explained 83%, and 8 PCs explained 92%. In subsequent experiments three versions were used: the first three PCs (***ProtFP_PCA (3)***), the first 5 PCs (***ProtFP_PCA (5)***) or all eight PCs (***ProtFP_PCA (8)***). See **Table 3.2** for the final principal components.

Table 3.2. Principal Components Resulting from the AAindex selection.

Amino Acid	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	Feature
Variance Explained	0.43	0.24	0.08	0.06	0.04	0.03	0.03	0.02	n/a
Total Variance Explained	0.43	0.67	0.75	0.81	0.85	0.88	0.90	0.92	n/a
G	-5.70	-8.72	4.18	-1.35	-0.31	2.91	0.32	-0.11	-176196525
A	-0.10	-4.94	-2.13	1.70	-0.39	1.06	-1.39	0.97	1169372512
C	4.62	-3.54	1.50	-1.26	3.27	-0.34	-0.47	-0.23	892384356
V	5.04	-2.90	-2.29	1.38	0.06	0.08	1.79	-0.38	-58134849
L	5.76	-1.33	-1.71	0.63	-1.70	0.71	-0.05	-0.51	-590269326
I	6.58	-1.73	-2.49	1.09	-0.34	-0.28	1.97	-0.92	-1784790725
M	5.11	0.19	-1.02	0.15	0.13	-0.30	-2.95	0.50	-188476976
F	6.76	0.88	0.89	-1.12	-0.49	-0.55	-0.87	1.05	-1561345091
W	7.33	4.55	2.77	-2.41	-1.08	1.04	0.23	0.59	-816166777
Y	3.14	3.59	2.45	-1.27	-0.06	-0.29	1.99	0.30	1237879003
H	0.17	2.14	1.20	0.71	1.16	-0.38	-1.85	-2.79	-1970548995
T	-2.00	-1.77	-0.70	1.02	1.06	-1.20	0.74	1.65	-266397547
P	-3.82	-2.31	3.45	1.00	-3.22	-3.54	-0.36	-0.30	-576206913
S	-4.57	-2.55	-0.67	1.11	0.99	-1.02	0.11	0.65	-1481898440
D	-6.61	0.94	-3.04	-4.58	0.48	-1.31	0.10	0.94	1957532765
N	-4.88	0.81	0.14	-0.14	1.23	-0.65	1.02	-1.94	-1593568836
E	-5.10	2.20	-3.59	-2.26	-2.14	1.35	-0.45	-1.31	558044215
Q	-3.95	2.88	-0.83	0.52	0.90	0.55	-0.08	0.64	-1986194934
K	-4.99	5.00	0.70	3.00	-1.23	1.41	0.19	0.87	268201585
R	-2.79	6.60	1.21	2.07	1.67	0.76	0.00	0.32	1636879004

Shown are all eight principal components and the variance explained by these principal components. In addition, the features obtained from the hashing of the AAindex selection are shown. This column represents the feature based ProtFP. Not available is abbreviated by n/a.

3.3.11 Distance between descriptor sets. To compare the characteristics of different descriptor sets and their behavior in describing particular AAs as similar and dissimilar, the average 'difference in distances' was calculated for each possible pair of descriptor sets. (See **Figure 3.1** for a scheme of the performed calculations). This value was obtained as follows. Firstly, a full similarity matrix was calculated for each possible AA pair using each descriptor set, thus consisting of 20*20 fields per descriptor set. The distances in this matrix were scaled linearly to a range between 0 (most similar) and 1 (most dissimilar).

Subsequently, for each possible pair of *descriptor sets* the *difference* between the *Euclidian distances of each AA pair* was calculated, giving rise to a total of 400 inter-amino acid distance differences per descriptor set pair. (In other words, we evaluated how differently two descriptors judged the difference between two AAs.

Given that 20 AAs exist, 400 distances exist between all AAs, for a single descriptor set – and the same number of *differences* of those distances for each descriptor set pair.) Of the 400 distances obtained, the average distance and the standard deviation was calculated and subsequently employed as a measure for the distance between amino acid descriptor sets (*i.e.*, if the average distance is high, two amino acid descriptor sets perceive similarities between amino acids in a very different way). The more different those distances are for different descriptor sets, the more different the particular descriptor sets considered behave. We employed a total of 12 descriptor sets for this amino acid descriptor comparison, since the feature based ProtFP descriptor set (***ProtFP (Feature)***) merely uses presence or absence of features and hence could not be included in the distance calculation. In the end, a matrix of 12*12 distances between descriptor sets was obtained which was subject to PCA with the aim to visualize the individual distances between descriptor sets in a graphical way.

(Conceptually, this work is similar to an analysis of chemical descriptors from the ligand side which was performed previously and given the importance of also comparing descriptors from the protein side the current work hence complements this study [30]).

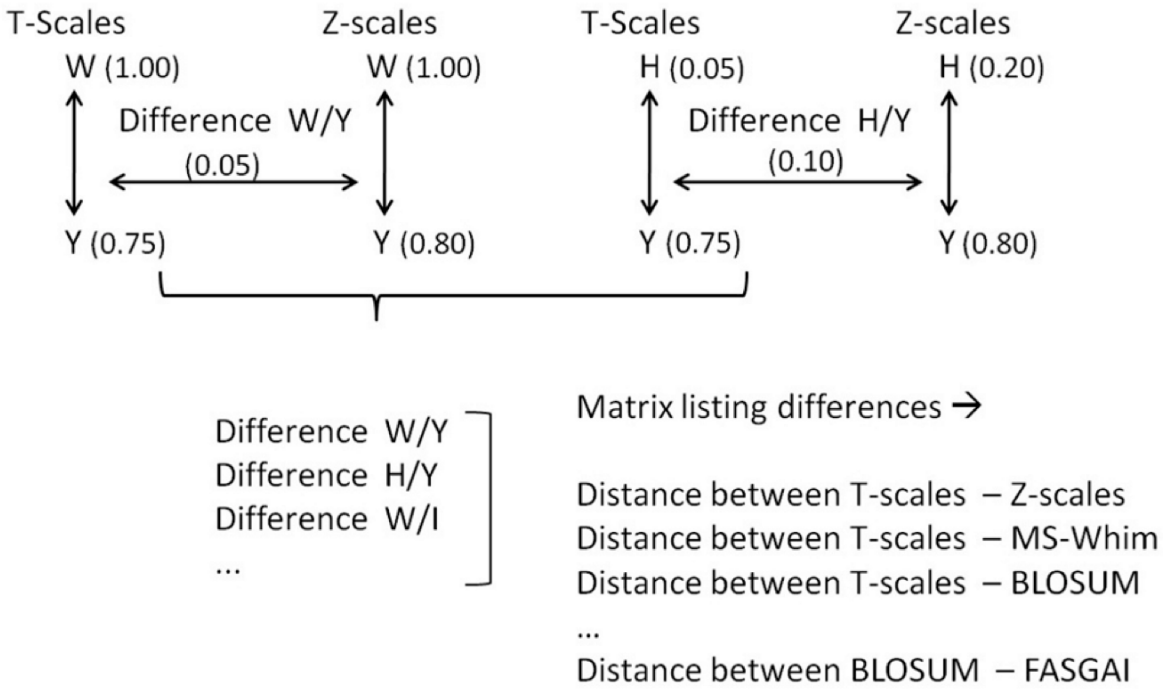


Figure 3.1: The approach used to characterize descriptor set distances and similarities. After normalization of all descriptor sets, the difference between a pair of descriptor sets was calculated. This difference was obtained as the difference between the distance separating pair of AAs in descriptor 1 and the same pair in descriptor 2. This was done for all descriptor set pairs. Finally, the average difference was obtained and a full matrix was constructed.

3.3.12 Benchmark datasets for different descriptors. While analyzing similar and different behavior of AA descriptor sets is relevant to judge *how similarly* two descriptor sets behave, it does not yet give any information how relevant the information captured by a particular descriptor would be for the generation of bioactivity models. Hence, in order to assess the performance of each descriptor set, three different data sets were used to perform a number of benchmark experiments.

ACE inhibitor data set. The first set consisted of 58 dipeptides with a measured ACE inhibiting effect (pIC_{50}) and was obtained from literature.[22] The set serves as a benchmark as several of the descriptor sets analyzed here were applied to this set in their original publication. Hence, it can demonstrate that the method we use (Random Forest) performs *on par* or better than the PLS which is conventionally used in QSAM publications (see Supporting **Table S6** for the comparison). See **Table 3.3** for further details about the data set.

Table 3.3. The Data Sets Used for the Bioactivity Benchmarks.

	ACE Inhibitors	**GPCRs**	**NNRTIs**
Total Size (Data Points)	58	4,951	4,024
Total Compounds	n/a	3,088	451
Average Compound Tanimoto Distance (ECFP_6)	n/a	0.89	0.02
Average Euclidian Distance Compounds (Physicochemical)	n/a	1.31	n/a
Total Targets (Peptides / Proteins)	58	26	14
Average Target Tanimoto Distance (ProtFP (Feature))	0.83	0.26	0.14
Average Euclidian Distance Target (ProtFP_PCA (3))	1.35	0.89	0.47
Completeness (% of total compound - target pairs)	-	0.06	0.64

GPCR data set. The first bioactivity data set employed for benchmarking different amino acid descriptors in PCM modeling comprised a subset of 26 human monoamine receptors (class A GPCRs listed in Supporting **Table S1**; see also Supporting **Figure S1** regarding the subset of receptors used) obtained from ChEMBL version 11.[23] Receptors were selected only if more than 120 unique ligands with annotated activity were present in ChEMBL. The trans-membrane (TM) binding site was defined according to Gloriam *et al.* and all residues selected were subsequently subject to conversion into numerical values using all protein descriptor sets listed above.[31]

For each of the 26 receptors included in this study all small molecules with an affinity on this receptor available in ChEMBL were selected and further narrowed down to only include Ki annotations with high confidence score (9). Compounds were then classified as 'active' (pKi > 7) or 'inactive' (pKi =< 7). Finally compounds were clustered (using the ECFP_6 fingerprint used to train the models) to obtain a total of 100 chemically diverse 'actives' and 100 chemically diverse 'inactives' per receptor. Compounds were standardized and ionized at pH 7.4 in Pipeline Pilot 8.5.[32]

In total 3,088 distinct compounds were selected to generate a bioactivity model, including 1,863 compounds with measurements on multiple GPCRs, hence leading to a final dataset comprising 4,951 ligand-protein data points (corresponding to 6 % of the total of 80,288 possible compound – receptor combinations in the full matrix of 3,088 compounds and 26 targets; see **Table 3.3** for further details)

NNRTI data set. The second bioactivity data set where PCM modeling was applied comprised of 14 mutants of HIV Non-Nucleotide Reverse Transcriptase Inhibitors (NNRTIs) and 451 compounds, hence a total of 6,314 possible compound – receptor combinations out of which for 4,024 a pEC_{50} value was available (66% of the total).[21] The compounds in this case were structural analogues, and hence (as opposed to the GPCR case) the average similarity between the compounds was high, as was the similarity between the protein targets since those were HIV mutants carrying 1 to 13 point mutations. Like in our previous work, the binding site was defined as those AAs that differed between the different mutants (a total of 24 residues).[21] The HXB2 / IIIB reference strain was defined as the wild type (See **Table 3.3** for further details).[33]

3.3.13 Amino Acid descriptor set benchmarking. Two different approaches were pursued to benchmark AA descriptor sets with respect to their ability to generate bioactivity models (and hence, to capture protein information relevant to bioactivity and ligand binding); namely 70-30 validation and Leave-One-Sequence-Out (LOSO) which are described in the following.

70-30 validation. The primary benchmark was a 70-30 validation experiment. Each descriptor set was used in turn in combination with each of the datasets, and a model was trained on a random 70% of the data available and used to predict the bioactivities of the remaining 30% of the data. This procedure was repeated three times and from the resulting validation parameters the average and standard error of the mean (SEM) was calculated. For the ACE inhibitors this represented a particularly challenging benchmark as this set only includes peptides and no small molecules. For the bioactivity datasets employed for PCM modeling, since these data sets include both proteins and small molecules, this benchmark provides an answer to two different questions.

Firstly, the model was asked to make bioactivity predictions for those compounds that are not present in the training set and hence to extrapolate in the *chemical domain*. This part of the validation was particularly emphasized in case of the GPCR data set due to the low average compound similarity. Hence the model is asked to extrapolate the activity of known compounds and targets to *unknown compounds*.

Secondly, a compound can be present in the training set as annotated on one target, and also be present in the test set as annotated on target 2. This part of the validation was hence emphasized in case of the NNRTI data set due to the high average compound similarity. In this case the model is asked to extrapolate the activity of known compounds and targets to *unknown combinations of the two*, while, individually, each chemical structure and sequence have been seen by the model before (but just not in this particular combination).

Leave-one-sequence-out validation. This validation experiment was performed for each *target* in order to assess extrapolation abilities of the PCM models in the biological / target domain. Hence this validation was only applied to the datasets containing targets (GPCR set and NNRTI set). This step is analogous to leaving out ligands from a dataset in more conventional bioactivity modeling – however, since PCM models are also able to extrapolate in the *biological domain* we also need to perform this additional validation here.

In this part of the work, repetitively a single target is left out of the training set and subsequently a model is trained on all bioactivity data points, except for those of the target in the test set, that was left out. Afterwards activity values of all compounds on the target left out of the initial training procedure are predicted and compared to the experimental values.

Again this procedure is repeated three times and the average and SEM were calculated of the validation parameters. These steps are repeated for all targets in the data set in turn. This type of validation is a specialty of PCM modeling since it takes advantage of its ability to extrapolate *also in target space*. It resembles both the real-world situation of deorphanizing receptors, taking only information from related proteins into account and attempting to identify bioactive chemical matter for a receptor for which no ligands have been identified yet.[34, 35]

(Also this concept is applicable to predict which drug to use against a particular receptor mutant in case of *e.g*. personalized medicines, such as in case of the question which drug to use against a particular HIV patient genotype which is addressed also in this work.) Since the ACE inhibitor set consisted of bioactive compounds only, LOSO could not be performed on this set.

3.3.14 Compound Descriptors. Ligands were described using ECFP_6 circular fingerprints,[36] which take into account the number of connections to an atom, the element type, the charge, and the atomic mass. These descriptors have previously been shown to perform well in comparative virtual screening studies.[30] This ligand side descriptor was employed for all studies presented in this work containing small molecules. Here an array size of 512 bits (each bit corresponding to a chemical substructure) was used.

In addition, in the GPCR data set compounds were described by their physicochemical properties. These properties were binned into classes, when compounds met one of these classes the property was set as '1', when they did not the property was set as '0' (analog to the Z-scales (Binned) descriptor). The classes that were used are available in Supporting **Table S4** and **S5**.

3.3.15 PCM Modeling Method. Both regression and classification models were generated in Pipeline Pilot Version 8.5 using the R-statistics modeling package version 2.12.1.[29, 32] Modeling was performed using the 'forest' package in R Statistics.[37] The size of the forest was experimentally determined to be optimal at 500 trees, the maximum number of descriptors allowed for each tree was set at a fraction 0.5 of the total number. Class size equalization was turned on and a performance estimate during training was obtained using out-of-bag validation. Furthermore data points were fed into the model in a randomized order (differing between repeats of an experiment) to get a more reliable performance estimate.

3.3.16 Model validation. To validate our models different parameters were employed depending on the modeling type. In regression models both the Root-Mean-Square Error (RMSE) and the correlation coefficient intersecting the origin (R_0^2) were employed.[38] For the classification models the Matthews correlation coefficient (MCC) was used to estimate model performance because of its robustness and the fact that it incorporates both correct and false predictions.[39] However, because of the importance of models to actually retrieve active compounds, we employed model sensitivity as a second performance measure.

3.3.17 Y-Scrambling. To make sure that the models created were not based on chance correlations, Y-scrambling or permutation testing was performed. These studies were performed using the same setup as the benchmark experiments (also in triplo) however the modeled variable (pIC_{50}, pEC_{50} or activity class) was randomized over the data points. Hence no correlation should exist between the descriptors (ligand and target) and the activity. The results are shown in Supplementary **Figures S36 – S40** and confirm that no predictive models can be trained on this randomized set.

3.3.18 Descriptor Ranking. Finally, to obtain a broadly derived performance measure we ranked all 13 amino acid descriptor sets based on their performance per dataset and experiment. This rank-based assessment prevents a single dataset that is modeled very well or very bad (as expressed in RMSE or MCC) unduly influencing the average performance of this descriptor set. Descriptor sets were ranked using the validation parameters (R_0^2 and RMSE in the case of regression and MCC and Sensitivity in the case of classification), the final rank per experiment is the sum of both validation ranks. For example in the ACE inhibitor set each descriptor would receive a rank based on the RMSE and one based on the $R_0^{2,}$ the final rank can hence be anywhere between 2 (best score on both validation parameters) and 26 (worst score on both validation parameters). Subsequently the descriptors were re-ranked between 1 and 13 to provide a final rank that could be compared over all three data sets.

3.4 Results and Discussion - Section 1 – Similarity between descriptor sets

The first part of our work covers the characterization of descriptor similarity between all benchmarked descriptor sets. Furthermore, we show how ProtFP based descriptor sets were derived.

3.4.1 PCA of final indices selection (ProtFP_PCA). Figure 3.2A shows the first two principle components of all 20 natural AAs when employing the ProtFP descriptor set. Overall, the plot shows a general clustering of AAs with similar properties with the first PC corresponding to hydrophobicity (F and I score high whereas D and E score low) and the second PC corresponding to size (W and K score high whereas G and A score low). Noteworthy is the clustering of Leucine and Isoleucine, which is intuitively correct due to their high chemical similarity, however not reproduced by all AA descriptors, like ST-scales (Supporting **Figure S17**). Furthermore, both charged (D, E and R, K) and aromatic residues (F, H, Y, W) cluster together. (The principle components, representing each AA in ProtFP space, can be found in **Table 3.2**.) Hence, overall the ProtFP descriptor set produces a clustering pattern that looks correct from a chemical point of view.

Figure 3.2B shows the loadings plot of the first 2 PCs that represent the ProtFP descriptor set. (For a complete list of indices used as input for the PCA please see Supporting **Table S2**.) Here, some interesting observations can be made. For instance, scale 24 and 43 correspond to AAindex FAUJ880112 and MONM990201, respectively. While the former is a measure for negative charge, the latter is a measure for 'averaged turn propensities in a transmembrane helix'. These two properties are close neighbors based on the first two components; however they have a relatively large distance in the third PC. This is interpretable in the following way: it is likely that charged residues, if present in a transmembrane region, initiates a turn and is therefore located at the edges of the TM region. Hence the clustering of these indices together can be rationally explained.

Scales 36 and 39 are another interesting case. The former corresponds to AAindex LEVM760102 (Distance between C-alpha and centroid of side chain) and the latter corresponds to LEVM760105 (Radius of gyration of side chain). It is interesting to see that these two indices end up so close together in the first, second and third principal component. However, this is indeed expected as the maximal range of gyration can only be large if the maximal distance possible between C-alpha and side chain center is large and vice versa.

In conclusion, the division of the AA over the principal component space seems interpretable and in agreement with biochemical intuition; this applies both to the scores and the loadings plot of the PCA we performed. The next step is to compare the new descriptor set ProtFP to existing descriptor sets that have previously been published, both with respect to their ability to capture similarities of AAs and their relative performance in incorporating protein information relevant to bioactivity into SAR models.

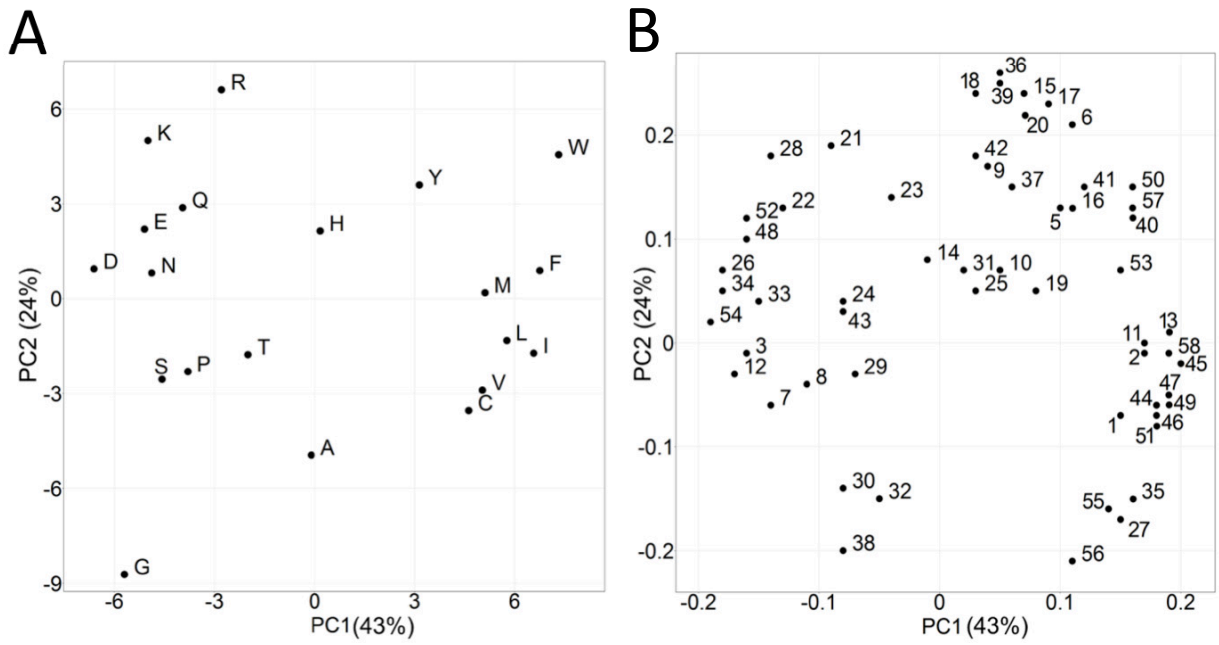


Figure 3.2: Principal components resulting from the PCA on 58 AAindices. (A) AAs that share physicochemical properties cluster together. The amount of variance explained by each principal component is shown in brackets. (B) The corresponding loadings plot where the numbers correspond to Supporting **Table S2**.

3.4.2 Distance between descriptors. Our first aim of the current study was to compare the behavior of AA descriptor sets, in order to investigate which descriptor sets agree on grouping AAs as similar, and which ones show largely orthogonal behavior. For this purpose, employing each of the AA descriptor sets a Euclidian distance based similarity matrix of all 20 by 20 AAs was calculated and visualized in a heat map. The comparison of ProtFP_PCA (3) with the frequently employed Z-scales (3) is shown in **Figure 3.3**. (The analogous plots, as well as numerical descriptions of the similarity matrices of other AA descriptor sets, are provided in Supporting **Tables S7** to **S18**, as well as Supporting **Figures S2** to **S13** for utilization by the reader in potential future studies).

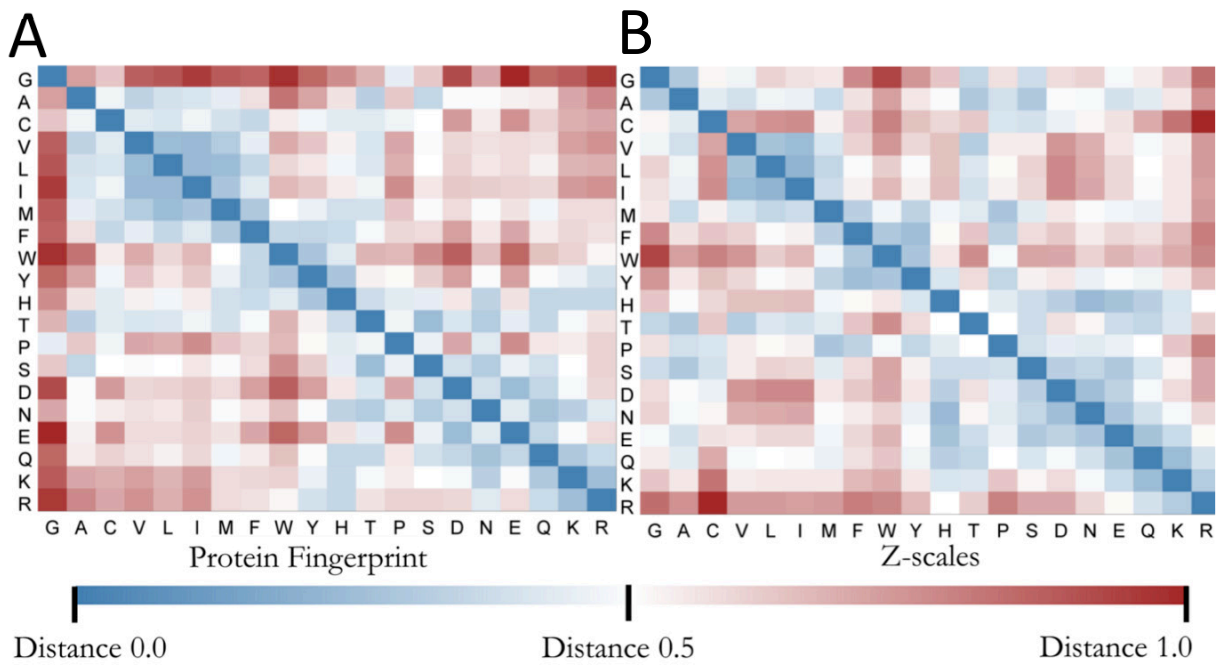


Figure 3.3: Comparison of the distances between individual AA pairs. (A) The heat map resulting from the ProtFP_PCA (3) similarity matrix. (B) The heat map resulting from the Z-scales analysis. In particular Histidine and Cysteine show a different distance spectrum when their similarity to the other AAs is compared.

Several clear differences are noteworthy when comparing the two descriptor sets. Firstly, the overall distances in the ProtFP_PCA (3) heat map are larger compared to Z-scales (3) despite the scaling that was applied. Furthermore, Glycine is located further away from the rest of the amino acids. Conversely, Cysteine is located closer to the aliphatic and aromatic AAs, but further away from the charged residues. Finally, Histidine also displays a different profile as it has a central position between the charged residues and aromatic residues in ProtFP PCA (3), whereas it is closely located to the charged AAs in Z-scales. Both descriptor sets therefore interpret the physicochemical space differently, while both views can be rationalized, benchmark experiments are needed to determine which leads to more predictive models.

As a next step it was considered how similar, on average, two descriptor sets perceive any pair of AAs, in order to establish how correlated their similarity perceptions are. **Figure 3.4A** shows the results of the PCA of the average distance between all descriptor sets, hence capturing the similarity in behavior of different AA descriptors. Shown are the 2 first PCs that explain 70 % of the variance. The first thing noteworthy is that MSWHIM, T-scales and ST-scales cluster together (here in the upper right quadrant); similarly, VHSE, FASGAI and ProtFP_PCA (3) form a second cluster (here in the lower right quadrant). The space between these two clusters is occupied by Z-scales (3) (upper right) and Z-scales (5) (lower right). ProtFP_PCA (5) and ProtFP_PCA (8) occupy the lower left quadrant but do not cluster. Finally Z-Scales (Binned) and BLOSUM behave distinctly from all descriptors above, and occupy the upper left quadrant. The distance between Z-scales (5) and Z-scales (Binned) is very large, which was not expected as one is constructed from the other. It could be speculated that the division into bins maximized separation between amino acids that only differ slightly on a continuous scale explaining the very different behavior. **Figure 3.4B** shows the results of the same PCA in three dimensions; now we observe that ProtFP_PCA (3) and Z-scales (5) are in addition to dissimilarities in the first two dimensions also out of the plane of the other descriptors.

The same calculation was repeated using only the absolute distance based on the first two PCs, comparing the descriptors based on the first two dimensions and minimizing the differences generated by a larger set of dimensions (Supporting **Tables S19 – S26** and **Figures S14 to S21**). Since we only use the first PCs the different versions of ProtFP_PCA are identical as are the versions of Z-scales. Again shown are the first two PCs which explain 66 % of the variance. Surprisingly, all descriptor sets based on a PCA of physicochemical properties form a cluster in this case (ProtFP PCA, VHSE and Z-scales), as do the two descriptors based on a topological description (T-scales and ST-scales). Contrarily, the MS-WHIM descriptor behaves most dissimilar to the others, likely due to the fact that this was the only descriptor constructed on an electrostatic potential. Finally, at first it seems surprising that the BLOSUM derived descriptor and the FASGAI descriptor are nearest neighbors in the first two principal components. However, in the 3rd principal component there is a large distance between the two points, rationalizing the difference.

Our results indicate that the different descriptor sets indeed describe the AA space differently, although there are commonalities most often based on the way they are constructed. What can be observed overall is that the use of more principal components (>3 per AA for a particular descriptor set) leads to a significant shift in the way they describe the AA differences.

This is true even while these principal components typically capture less variation of the original underlying matrix on which they were constructed. Therefore it stands to reason that the use of more than 3 principal components per AA might introduce less signal than noise (based on the small amount of variation captured by these components). Since the descriptor sets cluster mainly in the first 2 principal components of the descriptor analysis, these could be used as a guideline to determine complementarity when selecting descriptors to be used in bioactivity modeling (e.g. select one from each quadrant). Another observation is that the descriptor sets here introduced add novelty as they characterize the AA space differently. Assessing similarity in behavior is one aspect of comparing AA descriptor sets, in order to get an idea of the performance of the descriptor sets in this context we have set up several benchmark data sets as described in the following.

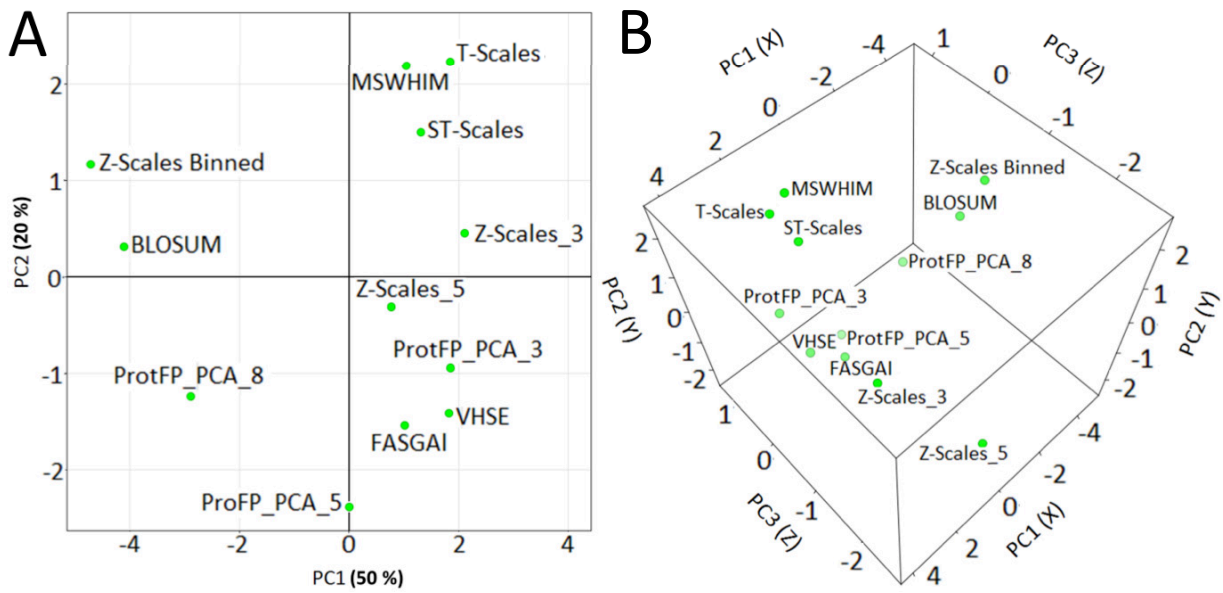


Figure 3.4: Principal component analysis of the distances between the different descriptor sets. Shown are the first two components (A). ProtFP_PCA (5) and (8) are seen to cluster away from the others. Furthermore T-scales, ST-scales and MSWHIM cluster together. (B) When the first three PCs are displayed Z-scales (5) and ProtFP_PCA (3) are seen to be distant from their cluster in the first two PCs.

3.5 Results and Discussion - Section 2 – Descriptor set performance in bioactivity models.

The second part of our work covers the assessment of descriptor set ability to create bioactivity models.

3.5.1 ACE inhibitors (70-30). The first benchmark we performed was a 70-30 validation experiment where ligands were dipeptides inhibiting ACE and where a random 70% of our data set was used for training and 30% for testing. The results of this validation on the test set are shown in **Figure 3.5.** The figure shows that all descriptor sets are capable of capturing the bioactivity space of the peptides as all have a RMSE under 0.8 log units. Interestingly, the best performing descriptor set is the Z-scales (Binned) descriptor (RMSE is 0.40 log units and the R_0^2 is 0.84), closely followed by the T-scales (RMSE 0.44 log units and R_0^2 0.86) and the Z-scales (3) (RMSE 0.41 log units and R_0^2 0.78). The worst performing descriptor set is ProtFP (Feature) (RMSE 0.74 log units and R_0^2 0.55), which is not surprising as it does not capture the different degrees of similarity between AAs, only that they are not the same. The ProtFP_PCA descriptor sets are performing better than ProtFP (Feature) but are still lagging compared with the other descriptor sets (RMSE approximately 0.10 log units higher and R_0^2 approximately 0.10 lower). The numerical values for the RMSE and R_0^2 are included as Supporting **Table S6**, also shown there are the training parameters Q^2 and cross validated RMSE (CV_RMSE) which are compared to values from previous studies for the same descriptors on the same set. We have constructed a PCA analysis of the similarity space formed by the dipeptides to explain the differences in behavior we observe. We hope to gain further insight in descriptor set performance by investigating how these descriptor sets characterize the different peptides.

3.5.2 ACE Inhibitors (Activity Space). We plotted the first two principal components for each descriptor set and colored the points by their pIC_{50} values (Supporting **Figure S22** – **S24**). We observe a direct correlation in the Z-scales (Binned) descriptor set between location in PCA space and activity. High affinity peptides score negatively on PC2, whereas all marginally active compounds score 0 or higher. Clearly the way the descriptors characterizes the peptides corresponds to their bioactivity. Conversely, the pattern obtained from the ProtFP (Feature) descriptor set does not clearly separate actives and inactives, explaining the poor performance. Well performing descriptor sets T-scales and Z-scales (3) and FASGAI also display a clustering similar to Z-scales (Binned). The PCA shows the highly active peptides to cluster together and the lesser actives are separated from these actives.

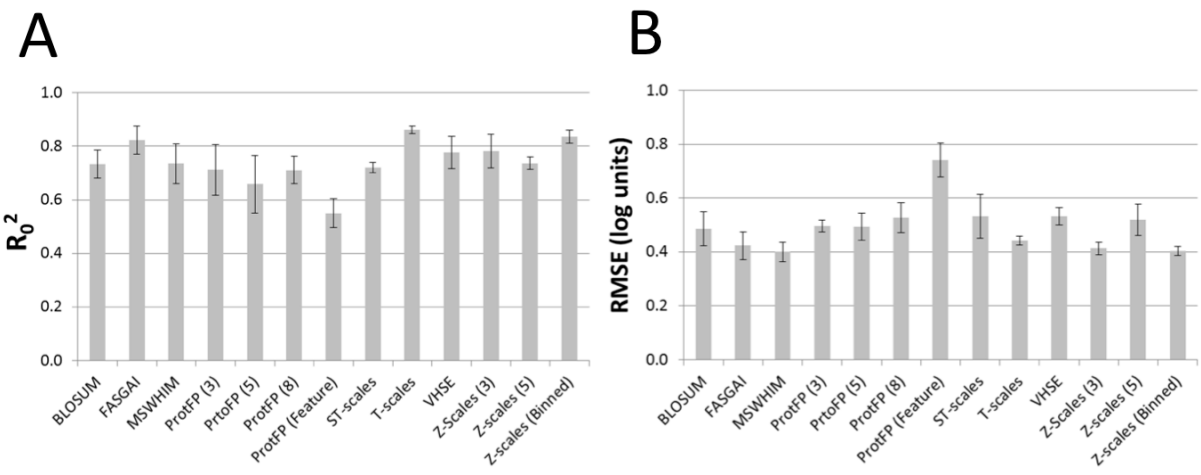


Figure 3.5: The average performance of the benchmarked descriptor sets in the ACE inhibitors 70-30 validation experiments. The average is calculated over three different experiments and the error bars represent the SEM. Shown are the R_0^2 (A) and the RMSE (B). While all descriptor sets perform similar, Z-scales (Binned) performs the best, followed by the T-scales, and ProtFP (Feature) performs the worst.

3.5.3 ACE inhibitors (Conclusions). We conclude that the differences in performance can be explained from the characterization of the peptides by each descriptor set as shown in the PCAs. In addition, we show that we can recreate models based on the individual descriptor sets that are comparable or better than previously published work. Finally, each descriptor set describes the AA space differently (as we have also shown in section 1). Still all were able to capture the bioactivity space and we therefore choose to apply these descriptor sets to PCM sets to see how well they perform.

3.5.4 GPCR ligands (70-30). Like we did with the ACE inhibitors, a similar 70-30 validation was performed on the GPCR set, although here a classification model was employed and performance was expressed as average sensitivity and MCC for all descriptors in the study (details are visualized in **Figure 3.6**). Here the descriptor sets perform much closer to each other compared to the ACE inhibitor set (all MCC values lie within the 0.35 – 0.40 range and all sensitivity values between 0.69 - 0.72), which is likely due to the much higher similarity of the targets and hence smaller space. Furthermore, the descriptor sets describe a smaller part of the space that is actually modeled since we now also include the chemical space next to the target space.

The best performance has been obtained in this case by the T-scales (MCC 0.39 and sensitivity 0.72), followed by Z-scales (5) (MCC 0.39 and sensitivity 0.71) and ProtFP_PCA (8) (MCC 0.39 and sensitivity 0.71). ProtFP (Feature) performs the worst (MCC 0.36 and sensitivity 0.69), but the difference is smaller than it was in the ACE inhibitor experiments. Another interesting observation is that all descriptor sets performed the best on the dopamine D5 receptor and the worst the histamine H3 receptor, irrespective of the protein descriptor set selected (Supporting **Figure S28**; although absolute differences in performance could be observed). These two receptors were also modeled the best and the worst respectively in the LOSO experiments as discussed in the following (where also a discussion of the likely underlying reason is given).

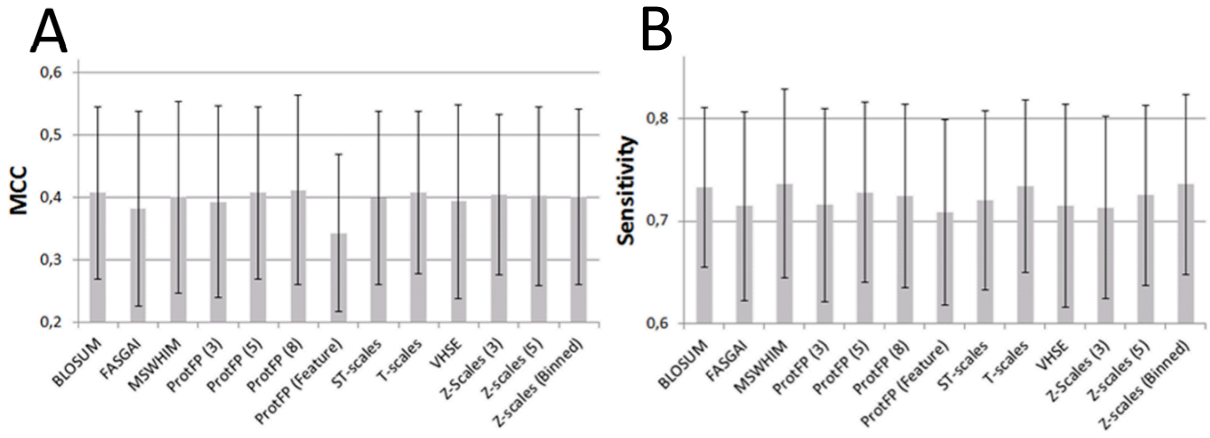


Figure 3.6: The average performance of the benchmarked descriptor sets in the GPCR 70-30 validation experiments. The average is calculated over all 26 receptors (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models, not between repeats of the individual models. Also see supporting **Figure S28**). Shown are the MCC (A) and the sensitivity (B). The differences between individual descriptor sets are smaller than in the ACE inhibitor experiments, likely due to the fact that models are based on both chemical and protein similarity. For individual receptors larger performance differences occur (main text). Still T-scales (3) performs the best and ProtFP (Feature) again performs the worst.

3.5.5 GPCR Ligands (LOSO). In order to benchmark the extrapolation capabilities of the descriptor set we performed a Leave-One-Sequence-Out experiment on the GPCR dataset, the results of which are shown in **Figure 3.7**. The overall performance is similar to the 70-30 validation but slightly worse (MCC between 0.29 – 0.32 and sensitivity between 0.57 – 0.60). However there are some differences, the best performance is by the Z-scales (3) (MCC 0.32 and sensitivity 0.59), followed by the ProtFP_PCA (5) (MCC 0.31 and sensitivity 0.60) and Z-scales (5) (MCC 0.32 and sensitivity 0.58). Surprisingly, the worst performance in this experiment is by ProtFP_PCA (8) (MCC 0.29 and sensitivity 0.57), yet it should be noted that the differences are marginal. Interestingly, the receptor that is modeled the best is again the dopamine D5 receptor and the worst the histamine H3 receptor, irrespective of the protein descriptor set selected (Supporting **Figure S29**). To gain a further understanding of this constant good performance for the D5 receptor and bad performance of the H3 receptor, we performed a PCA analysis analogously to the ACE inhibitors but then applied to the GPCR binding site sequences.

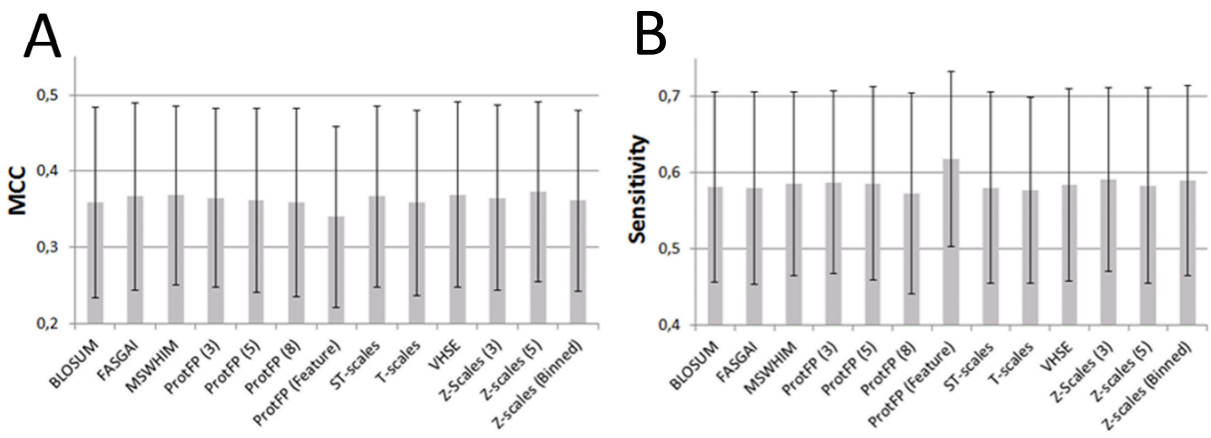


Figure 3.7: The average performance of the benchmarked descriptor sets in the GPCR LOSO validation experiments. The average is calculated over all 26 receptors (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different GPCRs, not between repeats of the individual models. Also see supporing **Figure S29**). Shown are the MCC (A) and the sensitivity (B). Here extrapolation takes place on the target side as the test set contains unseen targets. The differences between individual descriptor sets are still small. Again for individual receptors larger performance differences occur (main text). Now, Z-scales (3) performs the best and ProtFP_PCA (8) performs the worst.

3.5.6 GPCR Ligands (Target Space). From the PCA analysis of target space we can rationalize the poor performance on the histamine H3 receptor (Supporting **Figures S25** – **S27**). In the PCA of all GPCR targets used in this dataset, and employing the different descriptors, the H3 receptor is located at the edge of the PCA space. Furthermore, the three histamine receptors do not cluster together; in some cases the H3 receptor is located close to the H4 receptor, while in others it shows SAR that is closer to the H1 receptor. It is therefore likely that the models are unable to reliably extrapolate for this receptor based on the other two histamine receptors. Leaving out the H3 receptor removes crucial information from the SAR that cannot be compensated by the other two histamine receptors.

Conversely, the other receptor subtypes (5HT2, beta-adrenergic, and acetylcholine receptors) form clear sub-clusters, which hence allow leaving one receptor out while still retaining much information about the receptor space of that particular protein family. The well-performing dopamine D5 receptor on other hand is located at the center in all cases (always clustered with the other dopamine, 5HT1 and alpha-adrenergic receptors). Leaving this receptor out can therefore be considered straightforward as the target space is well covered

3.5.7 GPCR Ligands (Conclusions). We can conclude that all different descriptor sets can be used to create predictive PCM models on this set while still showing an order of (descending) performance as follows: Z-scales (5), ProtFP_PCA (3), T-scales, Z-scales (3), Z-scales (Binned), and BLOSUM (the latter two rank equal). The worst 3 are (descending): FASGAI, ST-scales, and ProtFP (Feature). Furthermore we can conclude that the binding site definition used for the GPCR descriptors is not optimal for all receptors. While the dopamine, 5HT1 and alpha-adrenergic are modeled very well, the histamine receptors clearly suffer, it would therefore be advisable to model these receptors with a different binding site definition. A starting point could be the work by Surgand et al. that also formed the basis for the paper by Gloriam *et al.* however Surgand *et al.* distinguish based on receptor family where Gloriam *et al.* produce a global selection.[40]

3.5.8 NNRTIs (70-30). While the above GPCR ligand dataset was based on rather diverse ligands, the NNRTI dataset employed in this study covers a more neatly defined area of both chemical (ligand) space, as well as biological (target) space. The first step is again a 70-30 validation experiment to assess the ability of the different descriptor sets to capture the ligand – target interaction space. The results are shown in **Figure 3.8**. Similar to previous experiments on the GPCR set, the performance of the descriptor sets is very similar (RMSE in the range 0.43 – 0.47 and R_0^2 in the range 0.56 – 0.61). However, in this set the ProtFP (Feature) performs the best (RMSE 0.43 and R_0^2 0.61), followed by MSWHIM (RMSE 0.44 and R_0^2 0.61) and Z-scales (3) (RMSE 0.44 and R_0^2 0.60). The worst performance comes from (descending) VHSE (RMSE 0.45 and R_0^2 0.59), BLOSUM (RMSE 0.45 and R_0^2 0.58), and ProtFP_PCA (8) (RMSE 0.46 and R_0^2 0.56).

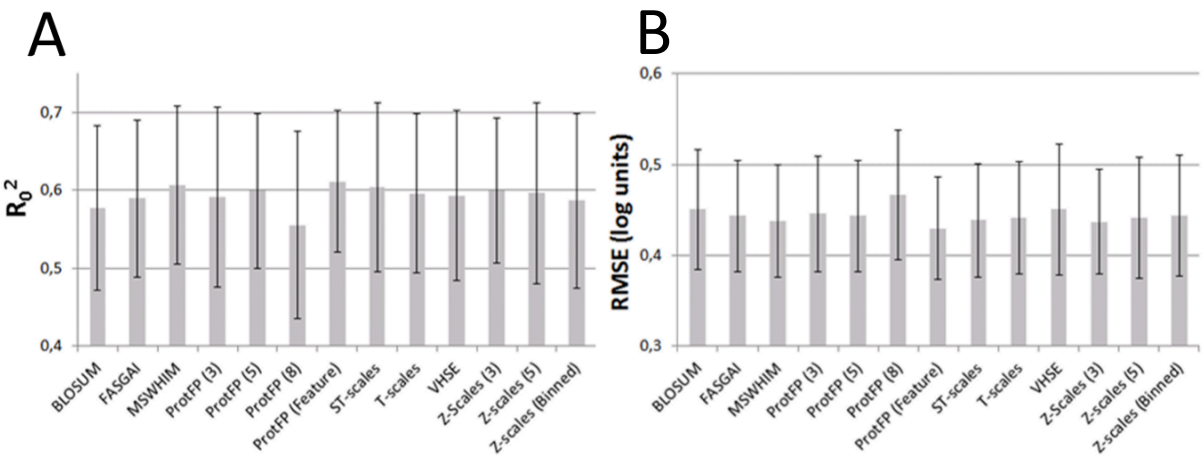


Figure 3.8: The average performance of the benchmarked descriptor sets in the NNRTIs 70-30 validation experiments. The average is calculated over all 14 mutants (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different mutants, not between repeats of the individual models. Also see supporting **Figure S30**). Shown are the R_0^2 (A) and the RMSE (B). Slightly more variance is seen compared to the GPCR experiments. In this case ProtFP_PCA (8) again performs the worst, while ProtFP (Feature) performs the best.

When focusing on the individual mutants (Supporting **Figure S30**), the best performing mutant is sequence 9 (carrying solely the K103N mutation, which is hence well covered in the remaining training set). All descriptor sets with the exception of BLOSUM are able to model the fraction of the compounds left out with an RMSE of < 0.3 log units on this mutant. The mutant that is modeled the worst is surprisingly not the heavy mutant sequence 7 (which contains a number of 13 total mutations), but rather sequence 2 carrying only two mutations (V179F and Y181C).

When comparing the predicted to the experimentally obtained bioactivities it can be seen that the bad performance is mainly caused by a number of outliers on the extremes. V179F is known to have a high impact on the class of compounds modeled here and the mutation itself (from valine to phenylalanine) is also a large change. Furthermore, this mutation was identified as having the most effect on binding in previous work.[21] The combination of these factors could explain the performance of all descriptors on this sequence. Still it should be noted that there are individual differences between descriptor sets (RMSE ranges between 0.54 – 0.66 and R_0^2 between 0.14 – 0.35). The next step we performed was to investigate whether results were transferable to the LOSO experiment, when extrapolation abilities to entirely novel sequences were required.

3.5.9 NNRTIs (LOSO). The LOSO validation was performed similar to the GPCR LOSO validation, leaving out one sequence at a time in training and predicting the activity of compounds on the sequence left out. The results are shown in **Figure 3.9**. The best performance is by BLOSUM (RMSE 0.73 and R_0^2 0.66), followed by ProtFP_PCA (3) (RMSE 0.73 and R_0^2 0.66) and ProtFP_PCA (5) (RMSE 0.73 and R_0^2 0.66), while the differences virtually absent. ProtFP (Feature) performs very well based on the RMSE (0.65), but based on the R_0^2 ranks 10th (0.64) and hence ranks 5th overall. The worst performance is obtained by (descending) Z-scales (3) (RMSE 0.75 and R_0^2 0.66), MSWHIM (RMSE 0.75 and R_0^2 0.64) and Z-scales (5) (RMSE 0.77 and R_0^2 0.64). Noteworthy is that, while the average RMSE rises to 0.7 log units, the average R_0^2 remains over 0.6 for all descriptor sets. This indicates that the descriptors are introducing an absolute error in the predictions, while still in most cases being able to accurately rank the compounds relative to each other.

The mutant modeled the best was sequence 3 (carrying only the Y181C mutation which is present multiple times in the data set, Supporting **Figure S31**). The sequence modeled the worst was sequence 8 (carrying K101P). This sequence was also modeled the worst in previous work.[21] The cause is likely that this particular mutation only occurs in sequence 7 and 8. Since sequence 7 is a heavy mutant, the model is unable to deconvolute the contribution of K101P tot the total effect on lowered binding of inhibitors. It is striking that ProtFP (Feature) performs so much better on this data set than the other two sets. On the NNRTI set, ProtFP (Feature) ranks 1st in the 70-30 validation and 5th in the LOSO validation, in the ACE inhibitor set it ranks 13th and the GPCR set 13th (70-30) and 9th (LOSO). To connect these observations of descriptor set performance to the similarity of the sequences and the way the descriptor sets characterize the space, we again performed a PCA analysis.

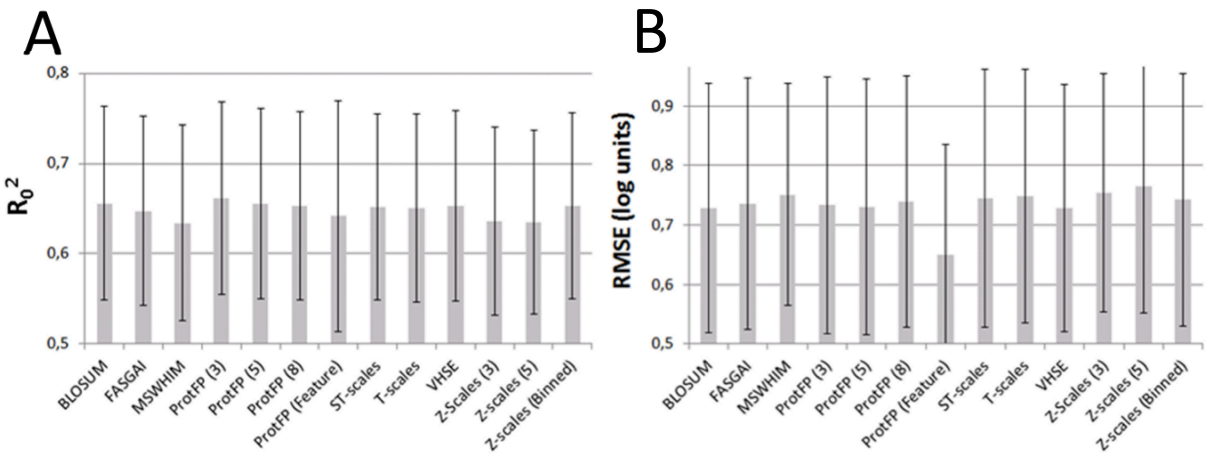


Figure 3.9: The average performance of the benchmarked descriptor sets in the NNRTIs LOSO validation experiments. The average is calculated over all 14 mutants (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different mutants, not between repeats of the individual models. Also see supporting **Figure S31**). Shown are the R_0^2 (A) and the RMSE (B). Here extrapolation takes place on the target side as the test set contains unseen targets. The differences between individual descriptor sets are still small but the spread of the SEM increases. Again for individual receptors larger performance differences occur (main text). Still ProtFP (Feature) again performs very good, it seems that a simplified representation is favorable for this data set.

3.5.10 NNRTIs (Target Space). The PCA analysis can explain the better performance of ProtFP (Feature) (Supporting **Figures S32 – S34**). Due to the fact that the mutants only differ by point mutations and one of the sequences caries 15 mutations (sequence 7), this sequence is set far apart from the other sequences by most descriptor sets. This effect is much less pronounced in ProtFP (Feature) as is does not differentiate between the type of mutations (all AAs are encoded as features so every amino acid difference is equal). The effect is that all the sequences cluster much closer than in the other descriptors, this leads to a better performance on this set.

Another cause for the observed effect could be that, by leaving out the residues that did not mutate in any of the sequences, we have maximized the dissimilarities to an extent that they do not accurately represent the bioactivity space. As the ProtFP (Feature) descriptor set leads to relatively small distances by merely encoding presence or absence of a feature, it partially compensates for this effect. In any case, the completely different way of describing the sequence similarity by ProtFP (Feature) proves to be beneficial for this dataset.

3.5.11 NNRTIs (Conclusions). The NNRTI set represented a different data set compared to the GPCR ligand dataset evaluated above as it consists of a number of highly similar sequences and compounds and, hence, resembles a typical data set one might encounter in lead optimization. We conclude that in these cases the feature base descriptor set might perform very well, however its good performance can also be catalyzed by the binding site definition. Therefore this type of descriptor set should be included as a possible candidate when working on a data set consisting of several highly related targets. For example a single GPCR subfamily like the adenosine receptors can also be considered a set of highly similar targets. The findings from this part could therefore also apply to this family. Indeed we found in other work that ProtFP (Feature) also performs well on this set.[20]

3.5.12 Final Descriptor Set Ranking. The final ranking of the individual descriptor sets is given in **Table 3.4**. This table included the individual ranks of all descriptor set in each experiment (on a scale of 1 to 13) and a final overall ranking (the sum of the individual rankings). We have also included the average rank and the SEM of this average rank (**Figure 3.10**).

The best performing descriptor sets overall are T-scales (3) (average rank 5.2), ProtFP_PCA (3) (average rank 5.4), Z-scales (3) (average rank 5.6) and Z-scales (Binned) (average rank 6.0). Taking into account that all descriptor sets performed very close and the easy interpretability from the Z-scales (Binned) might make this descriptor set the best choice to use for PCM experiments (it also displays the smallest SEM of the four).

The worst performing descriptor sets are ProtFP (Feature) (average rank 8.4), ST-scales (average rank 9.0), and ProtFP_PCA (8) (average rank 9.4). While their performance was close to the other descriptor sets, they were in the lower performing ranks in 80 % of the experiments. Therefore it might be wise to avoid these descriptor sets on bioactivity modeling in setups such as the PCM modeling employed here; but this again will surely depend on the particular dataset at hand as well.

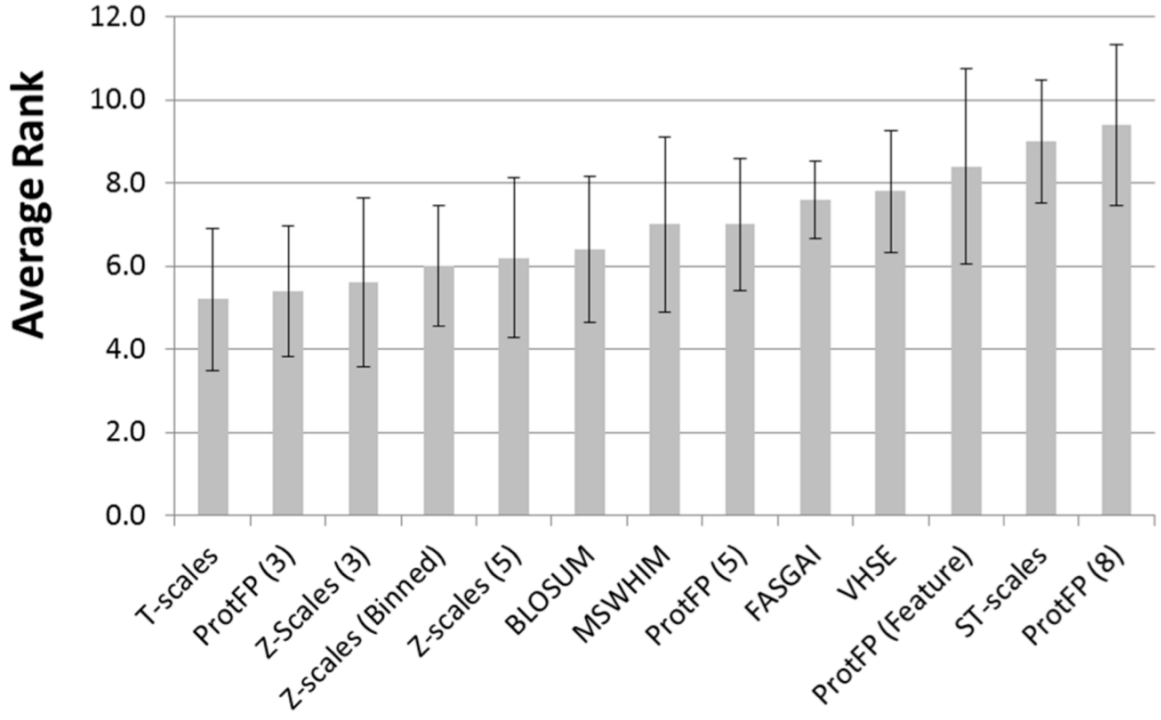


Figure 3.10: The average rank of the descriptor sets in the bioactivity benchmarks. The average is calculated over these 5 ranks and the SEM is given by the error bars. The best three descriptor sets perform about equal with an average rank =< 6 (where Z-scales (Binned) shows the smallest spread). The worst performance is by ProtFP (Feature), ST-scales and ProtFP_PCA (8) with an average rank > 8. ProtFP (Feature) and MSWHIM have a large error bar due to their inconsistent performance.

Contrary to our expectations, the best performing descriptor set is based on 135 amino acids (including non-natural amino acids) rather than one built on only the natural amino acids. Another interesting observation is that the top 4 descriptors on average consist on 4 principal components and the worst 4 consist of on average 8 principal components. Furthermore when multiple versions of the same descriptor set are compared, the set employing the least number of amino acids consistently scores better (ProtFP_PCA and Z-scales). This confirms our expectation that including more principal components that describe less variance introduces more noise than information.

Table 3.4. Overall Descriptor Set Ranking.

Descriptor	Final Rank	Rank ACE Inhibitors	Rank GPCR 70-30 validation	Rank GPCR LOSO	Rank NNRTI 70-30 validation	Rank NNRTI LOSO	Mean Rank
T-scales	26	2	1	8	5	10	5.2 (±1.7)
ProtFP (3)	27	9	5	2	9	2	5.4 (±1.6)
Z-Scales (3)	28	3	10	1	3	11	5.6 (±2.0)
Z-scales (Binned)	30	1	6	7	10	6	6.0 (±1.5)
Z-scales (5)	31	7	2	3	6	13	6.2 (±2.0)
BLOSUM	32	6	7	6	12	1	6.4 (±1.8)
MSWHIM	35	5	12	4	2	12	7.0 (±2.1)
ProtFP (5)	35	10	4	11	7	3	7.0 (±1.6)
FASGAI	38	4	9	9	8	8	7.6 (±1.0)
VHSE	39	8	11	5	11	4	7.8 (±1.5)
ProtFP (Feature)	42	13	13	10	1	5	8.4 (±2.4)
ST-scales	45	12	8	12	4	9	9.0 (±1.5)
ProtFP (8)	47	11	3	13	13	7	9.4 (±2.0)

The descriptor sets are sorted based on their final rank. Also shown are the rank each descriptor set receives in each individual benchmark. The final column shows the average rank for each descriptor set (calculated from the 5 individual ranks) and the SEM of associated with this average. The best performance is achieved by the T-scales (3), closely followed by ProtFP (3), Z-scales (3) and Z-scales (Binned).

3.5.13 Training Times. One final property of the descriptor sets has not been highlighted yet. On a workstation with a core i7 860 CPU and 16 GB memory, we found considerable differences in training times for the individual descriptor sets. On the datasets used in this work, as a rule of thumb ProtFP Feature showed the fastest model training while BLOSUM required most time (191% of the training time required for ProtFP Feature). The reason for this large difference is that the feature based descriptor set uses a single variable per amino acid, where the numerical descriptor sets use 3 (ProtFP PCA (3), Z-scales (3) and MS-WHIM) to 10 values (BLOSUM).

3.6 Conclusions

Given the large number of AA descriptor sets available we aimed to both characterize those descriptor sets with respect to their perception of similarities between AAs, and to benchmark them in bioactivity models. Descriptor set clustering indicated that they show different behavior from one another when characterizing AA similarities. As might be intuitive, when only considering the first two principal components, descriptor sets cluster the way they are derived, with Z-scales, VHSE and ProtFP PCA falling into one cluster, T-scales and ST-scales forming a second group of descriptor sets, and FASGAI, BLOSUM and MS-WHIM descriptor sets being somewhat distinct to the above groups.

Our results confirm that all QSAM descriptor sets can be used to train predictive bioactivity, including PCM, models. Performance differences between descriptor sets were in the order of magnitude of an RMSE difference of 0.1 log units. Individual targets could cause much larger differences in performance (e.g. the RMSE difference between the HIV mutant modeled best and worst was 1.2 log units). Therefore we conclude that all descriptor sets can be used to create predictive models.

Nevertheless, depending on the problem at hand, it might be wise to do an initial descriptor set selection before training a final model. In particular in data sets where affinity on unknown targets is predicted (like receptor deorphanization exercises, simulated by our LOSO experiments), larger differences in performance can occur. In cases where virtual screening is applied to a data set consisting of known targets, these differences are slightly smaller in magnitude.

3.7 Acknowledgements

The financial support of Tibotec BVBA is gratefully acknowledged.

3.8 Supporting Information

Additional tables (Supporting **Tables S1 – S26**), figures (**Figures S1 – S39**) are available as pdf. Furthermore, we include a Pipeline Pilot component to convert single letter AA sequences to any of the here tested descriptor sets and a fully functional example protocol, both to be used in Pipeline Pilot 8.5 and up (archive file). These materials are available online at www.gjpvanwesten.nl. The GPCR data set is available upon request but was considered too large to submit with the paper.

3.9 References

1. A. Kontijevskis, P. Prusis, et al.; *A look inside HIV resistance through retroviral protease interaction maps.* PLoS Comput. Biol.; *2007*. **3** (3): e48.
2. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
3. G.J.P. Van Westen, J.K. Wegner, et al.; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.
4. J.E.S. Wikberg, F. Mutulis, et al.; *Melanocortin receptors: Ligands and proteochemometrics modeling;* in *Melanocortin System*; D. Braaten; Editor *2003*: New York. p. 21-26.
5. J.R. Bock and D.A. Gough; *Virtual screen for ligands of orphan G protein-coupled receptors.* J. Chem. Inf. Model.; *2005*. **45** (5): 1402-1414.
6. M. Lapinsh, P. Prusis, et al.; *Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.* Mol. Pharmacol.; *2002*. **61** (6): 1465-1475.
7. P. Prusis, S. Uhlén, et al.; *Prediction of indirect interactions in proteins.* BMC Bioinformatics; *2006*. **7**: 167-180.
8. J. Jonsson, T. Norberg, et al.; *Quantitative sequence-activity models (QSAM)--tools for sequence design.* Nucl. Acids Res.; *1993*. **21** (3): 733-739.
9. M. Sandberg, L. Eriksson, et al.; *New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids.* J. Med. Chem.; *1998*. **41** (14): 2481-2491.
10. J. Meslamani, J. Li, et al.; *Protein–Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling.* J. Chem. Inf. Model.; *2012*. **52** (4): 943-955.
11. H. Strombergsson, A. Kryshtafovych, et al.; *Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures.* Proteins: Struct., Funct., Bioinf.; *2006*. **65** (3): 568-579.
12. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.
13. P. Zhou, F. Tian, et al.; *Quantitative Sequence-Activity Model (QSAM): Applying QSAR Strategy to Model and Predict Bioactivity and Function of Peptides, Proteins and Nucleic Acids.* Current Computer - Aided Drug Design; *2008*. **4** (4): 311-321.

14. L. Yang, M. Shu, et al.; *ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues.* Amino Acids; *2010*. **38** (3): 805-816.

15. H. Mei, Z.H. Liao, et al.; *A new set of amino acid descriptors and its application in peptide QSARs.* Biopolymers; *2005*. **80** (6): 775-786.

16. F. Tian, P. Zhou, and Z. Li; *T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides.* J. Mol. Struct.; *2007*. **830** (1-3): 106-115.

17. L. Guizhao and L. Zhiliang; *Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides.* QSAR Comb. Sci.; *2007*. **26** (6): 754-763.

18. A. Zaliani and E. Gancia; *MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies.* J. Chem. Inf. Comput. Sci.; *1999*. **39** (3): 525-533.

19. A.G. Georgiev; *Interpretable numerical descriptors of amino acid space.* J. Comput. Biol.; *2009*. **16** (5): 703-723.

20. G.J.P. Van Westen, O.O. van den Hoven, et al.; *Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data.* J. Med. Chem.; *2012*: **55** (16): 7010-7020.

21. G.J.P. Van Westen, J.K. Wegner, et al.; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

22. S. Hellberg, L. Eriksson, et al.; *Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships.* Int. J. Pept. Protein Res.; *1991*. **37** (5): 414-424.

23. A. Gaulton, L.J. Bellis, et al.; *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res.; *2011*. **40**: D1100 - D1107.

24. S. Hellberg, M. Sjoestroem, et al.; *Peptide quantitative structure-activity relationships, a multivariate approach.* J. Med. Chem.; *1987*. **30** (7): 1126-1135.

25. M. Lapins, M. Eklund, et al.; *Proteochemometric modeling of HIV protease susceptibility.* BMC Bioinformatics; *2008*. **9** (1): 181-192.

26. M. Connolly; *Analytical molecular surface calculation.* J. Appl. Crystallogr.; *1983*. **16** (5): 548-558.

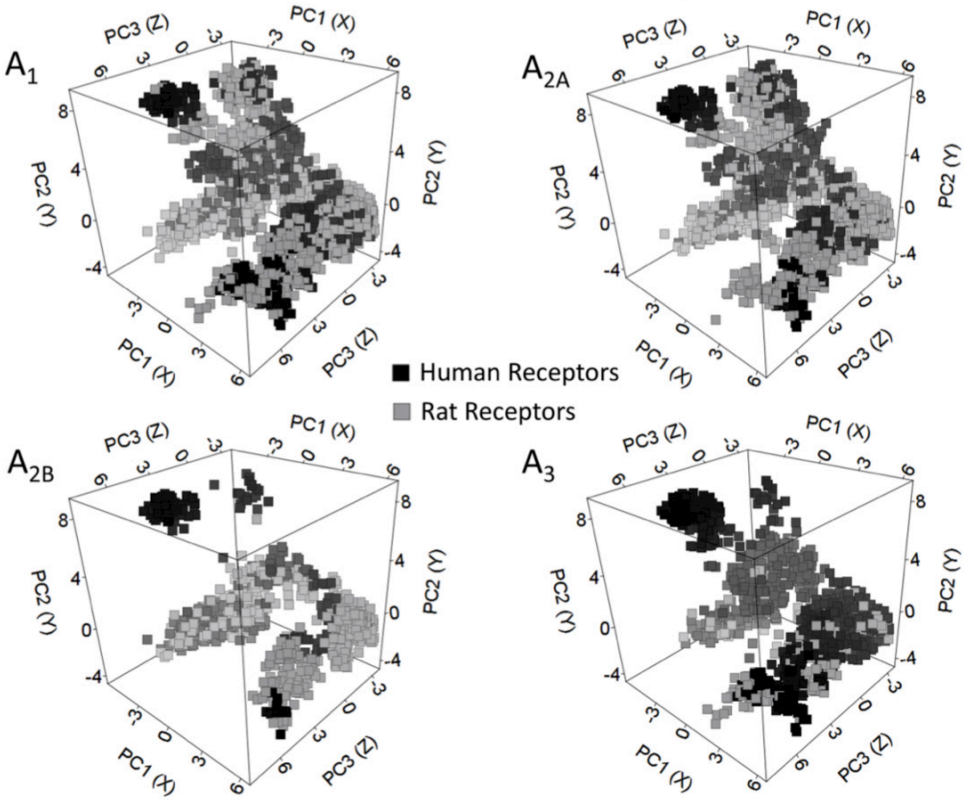
27. S. Henikoff and J.G. Henikoff; *Amino acid substitution matrices from protein blocks.* Proc. Natl. Acad. Sci. U. S. A.; *1992*. **89** (22): 10915-10919.

28. S. Kawashima, H. Ogata, and M. Kanehisa; *AAindex: Amino Acid Index Database.* Nucleic Acids Res.; *1999*. **27** (1): 368-369.

29. R Development Core Team; *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing *2006*.
30. A. Bender, J.L. Jenkins, et al.; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.
31. D.E. Gloriam, S.M. Foord, et al.; *Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design.* J. Med. Chem.; *2009*. **52** (14): 4429-4442.
32. Accelrys Software Inc *Pipeline Pilot Professional Edition* Scitegic Version 8.5
33. B.T. Korber, B.T. Foley, et al. *Numbering Positions in HIV Relative to HXB2CG*. *1998*.
34. E. van der Horst, J. Peironcely, et al.; *A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization.* Bmc Bioinformatics; *2010*. **11** (1): 316.
35. E. Van der Horst, J.E. Peironcely, et al.; *Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1964-1977.
36. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.
37. A. Liaw and M. Wiener; *Classification and Regression by randomForest.* R News; *2002*. **2** (3): 18-22.
38. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.
39. P. Baldi, S. Brunak, et al.; *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics; *2000*. **16** (5): 412-424.
40. J.-S. Surgand, J. Rodrigo, et al.; *A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **62** (2): 509-538.

Chapter 4

Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data



G.J.P. Van Westen, O.O. van den Hoven, R. van der Pijl, T. Mulder-Krieger, H. de Vries, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; J. Med. Chem.; 2012: ***55*** *(16): 7010-7020.*

Contents

4.1 Abstract

The four subtypes of adenosine receptors form relevant drug targets in the treatment of e.g., diabetes and Parkinson's disease. In the present study we aimed at finding novel small molecule ligands for these receptors using virtual screening approaches based on proteochemometric (PCM) modeling. We combined bioactivity data from all human and rat receptors in order to widen available chemical space. After training and validating a proteochemometric model on this combined dataset (Q^2 of 0.73, RMSE of 0.61) we virtually screened a vendor database of 100,910 compounds. Of 54 compounds purchased, six novel high affinity adenosine receptor ligands were confirmed experimentally, one of which displayed an affinity of 7 nM on the human adenosine A1 receptor. We conclude that the combination of rat and human data performs better than human data only. Furthermore, we conclude that proteochemometric modeling is an efficient method to quickly screen for novel bioactive compounds.

4.2 Introduction

4.2.1 The Adenosine Receptors. G protein-coupled receptors (GPCRs) are membrane-bound proteins and targets for many hormones and neurotransmitters in the body. As such they are ideal drug targets with a large degree of inherent selectivity due to their tissue specific expression. The local hormone adenosine interacts with four different GPCRs, the adenosine A_1, A_{2A}, A_{2B} and A_3 receptors. These receptor subtypes are involved in many (patho)physiological processes, including diseases such as type 2 diabetes, heart arrhythmias, and Parkinson's disease.[1] In the current work we set ouy to identify novel small molecule ligands for these adenosine receptors using virtual screening approaches.

4.2.2 Proteochemometric Modeling. Different approaches exist to select potentially bioactive compounds using computational models. Conventionally, a structure-activity model can be created using known compounds.[2, 3] The obtained model can then be used to predict the modeled output variable for compounds that have not been experimentally tested, on the basis of the 'Molecular Similarity Principle' which states that similar compounds show similar activity.[4] However, in the case of the adenosine receptors we are dealing with *multiple similar* targets, rather than one target. Previously it has been shown that proteochemometric modeling (PCM),[5-8] is able to create robust predictive models for multiple similar targets.[9-11] As has been reviewed in detail before,[7] PCM takes both ligand- as well as target-similarity into account, and can thereby also benefit from the principle that '*similar targets bind similar ligands*'. Given the ability of PCM models to also consider ligands active on related receptors when predicting bioactivity against a particular receptor, this increases the likelihood of identifying both active compounds and novel active chemotypes. Hence, we chose to create a PCM model trained on the adenosine receptor subfamily, rather than to train individual bioactivity models. We hypothesized the PCM model to perform better than these individual models and hoped to find both compounds that are a selective ligand for a single receptor but also compounds that are globally active ligands active on the entire subfamily of human adenosine receptors.

4.2.3 Chemical space and target space. Chemical space can be characterized based on the similarity of the compounds that interact with adenosine receptors. It is this space that is exploited when a structure-activity model is created for one of the receptors, as chemicals predicted to be closely located to known ligands on a target are expected to be ligands of that target. Target space can be characterized by the similarity of the targets. It is this space that is exploited when a multiple sequence alignment shows that the A_{2A} and A_{2B} adenosine receptors are more similar than the A_{2A} and A_3 receptors (Supporting **Table S1**).[12] Since PCM uses both ligand and target to predict an output variable, it can thereby also consider the fact that some features have different effects on different targets to better fit the data.[13, 14] Therefore our hypothesis in the current work is that PCM models perform better in prediction of values for data points that were not originally in the training set when compared to conventional structure-activity models, a principle that has been shown before for different mutants of HIV reverse transcriptase.[10]

4.2.4 Inclusion of multiple species orthologs. Historically rat tissues have been the source of adenosine receptors for the testing of novel chemical entities, before the human receptors became available for in vitro testing.[12] As a result a large amount of historical data is available on the rat orthologs (identical receptors in other species) of the human adenosine receptors including affinity data of small molecules (5,397 data points in ChEMBL 2).[15] Recently it has been shown that in general, small molecule binding is conserved between human and rat orthologs. However, a species specific pharmacology is observed for the A_1 and A_{2A} receptor; relative to human receptors the average pKi for the A_1 receptor is -0.51 log units while 0.41 log units for the A_{2A} receptor.[16] Moreover, from the full sequence similarity it becomes apparent that rat and human orthologs show greater similarity (identity 84% ± 8) than human paralogs (similar receptors in the same species) among each other (48% ± 7) (Supporting **Table S1**).

As it has been previously shown that PCM can model paralog subfamilies,[17, 18] in this work we extend this approach by proposing to include *orthologs* in the training set, in order to capture the chemical space associated with these orthologs in the model as well, while at the same time considering target differences in the PCM model generation process. Through a combined virtual and experimental screening we hope to find both novel selective ligands (active on a single receptor) and novel global ligands (active on all human adenosine receptors) as both these ligand types are of interest to our research team.

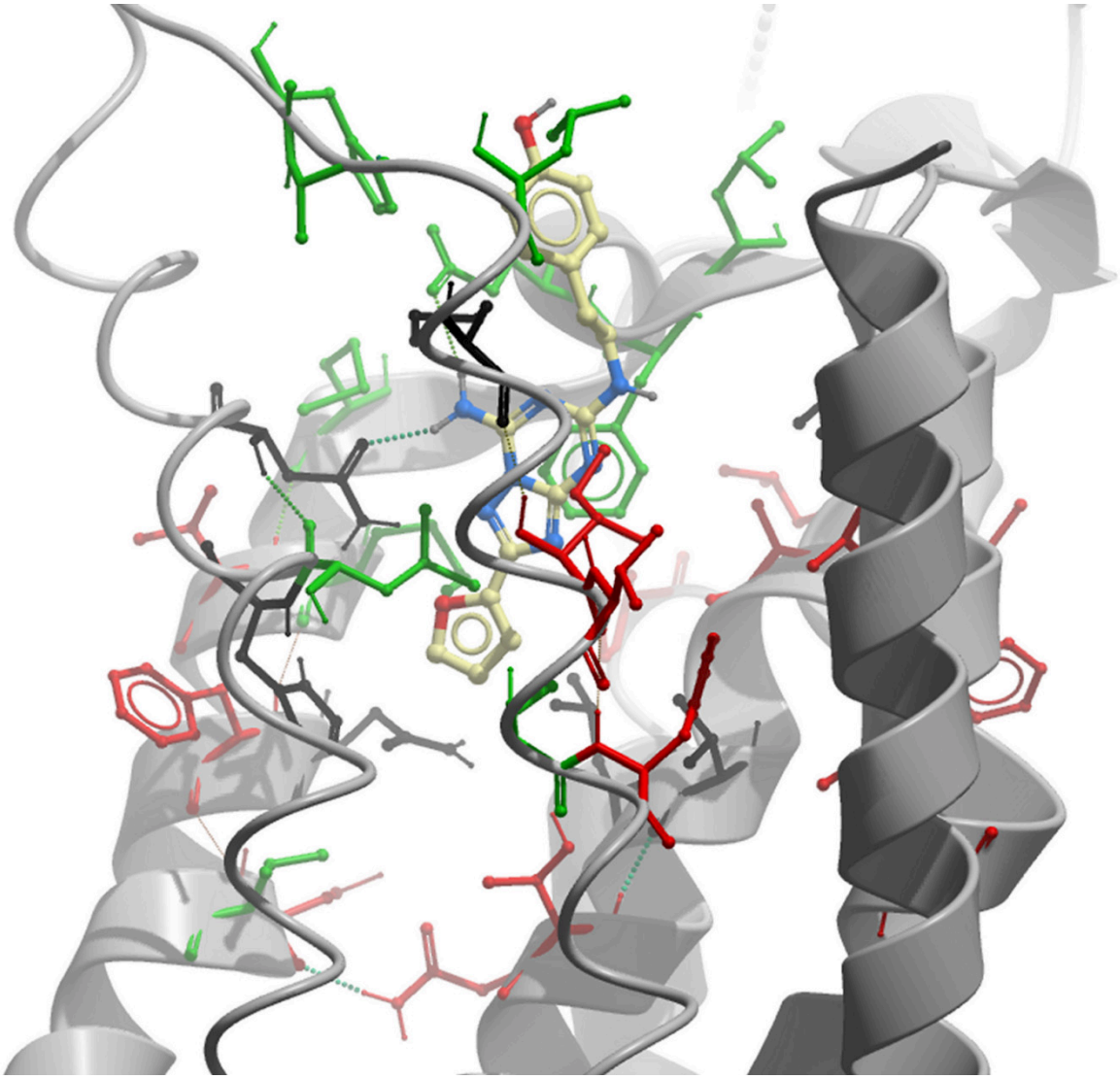


Figure 4.1: The binding site we selected to define the target similarity as visualized in PDB structure 3EML. The protein backbone is in gray and the co-crystallized ligand (ZM-241385) in ball and stick model. The green residues were obtained through selection of a 5 Å sphere around the co-crystallized ligand, red residues were obtained through TEA analysis (see **4.5.4**) and residues in gray/black occur in both analyses. Note that residues in both trans-membrane domains and extracellular loops were included.

4.3 Results and Discussion

4.3.1 Characterizing target space. Figure 4.1 shows the residues selected as binding site displayed in the adenosine A_{2A} receptor crystal structure containing antagonist ZM241385 (PDB code 3EML).[19] The figure displays the trans-membrane (TM) domains and extracellular loops (ELs) of the receptor. Individual amino acid side chains have been visualized in ball and stick model. The green residues were obtained through selection of a 5 Å sphere around the co-crystallized ligand, red residues were obtained through Two-Entropy Analysis (TEA) (see section **4.5.4**) and residues in gray/black occur in both analyses. **Figure 4.2** displays the results from a principal component analysis (PCA) of the ligand binding pocket in all receptors. The binding pocket was defined based on the same residues that were used to train the final model, representing *target space* (see section **4.5.4**). The PCA demonstrates that our binding site retains the pattern from full sequence similarity, in which receptor orthologs are more similar than paralogs. It should be noted however, that the difference between rat and human orthologs of the A_3 receptor is much larger than in any of the other three ortholog sets. This large difference is in agreement with the fact that compounds found to be active on the human A_3 receptor were much less active or even inactive on the rat A_3 receptor.[1] Therefore a full clustering of these two receptors based on the binding site would be contradictory to what we know from the chemical space of the two receptors, which is described below.

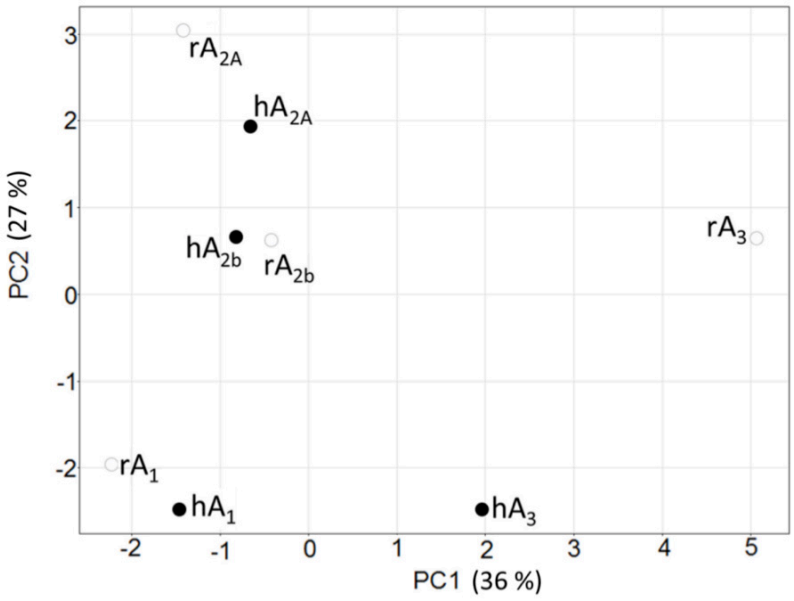


Figure 4.2: Principal component analysis of the similarity in target space. The adenosine receptor orthologs are more similar than their paralogs. human receptors are indicated with a black circle and rat receptors by a white circle. Both A_3 receptors are very different ('outliers'), while the A_{2A} and A_{2B} receptors cluster together. This observation is consistent with the fact that ligands active on the human A_3 receptor were often found to be inactive (or less active) on the rat A_3 receptor.

4.3.2 Characterizing chemical space. In addition to the analysis of target space, we also performed a PCA on the structures of all compounds (using the same descriptor as the final model) we had available in our data set, or *chemical space*, comprising a total of 10,999 data points. The results are displayed in **Figure 4.3** and data points are grouped by orthologs. In the same way as it could be observed in target space, a high similarity between orthologs is also visible in the chemical (ligand) space. Furthermore it becomes apparent that the chemical spaces for the A_1 and A_{2A} orthologs have been explored most extensively, while the chemical space for the A_{2B} orthologs has been sampled rather sparsely. Finally, the chemical space for the A_3 orthologs is dominated by compounds measured on the human ortholog, biasing in particular this dataset of active compounds. In fact, as mentioned earlier it has been hard to identify ligands (and in particular antagonists) that exhibit affinity for the rat A_3 receptor in previous work.[1]

The results from this PCA analysis show that there is significant (however in no case complete) overlap in the chemistry of the compounds that have been tested on the human as well as rat adenosine receptor subtypes (chemical space). Nevertheless, the number of identical compounds tested per ortholog pair is rather low (at about 5% of the total number of active compounds in our dataset, Supporting **Table S2**).

Likewise we compared the chemical space between the *paralogs* for both human and rat receptors (Supporting **Figure S1** and **S2**). The chemical space of annotated compounds for paralogs in the training set is very similar with the exception of the A_{2B} receptors. However, the points are colored according to their affinity on the receptors showing that the location of 'high affinity hotspots' differs between paralogs, while some hotspots are shared. This observation of high affinity hotspots' confirms that chemistry alone cannot explain the affinity differences between receptors, but also that selective compounds can be found within the training set, hence we expect our model to be able to predict selectivity.

The analysis of chemical space gave us confidence that bioactivity space between human and rat adenosine receptor orthologs is similar enough to allow us the use of PCM modeling approaches; still that it is also dissimilar enough to enable the discovery of novel bioactive ligands by considering bioactive space from both species in a single model.

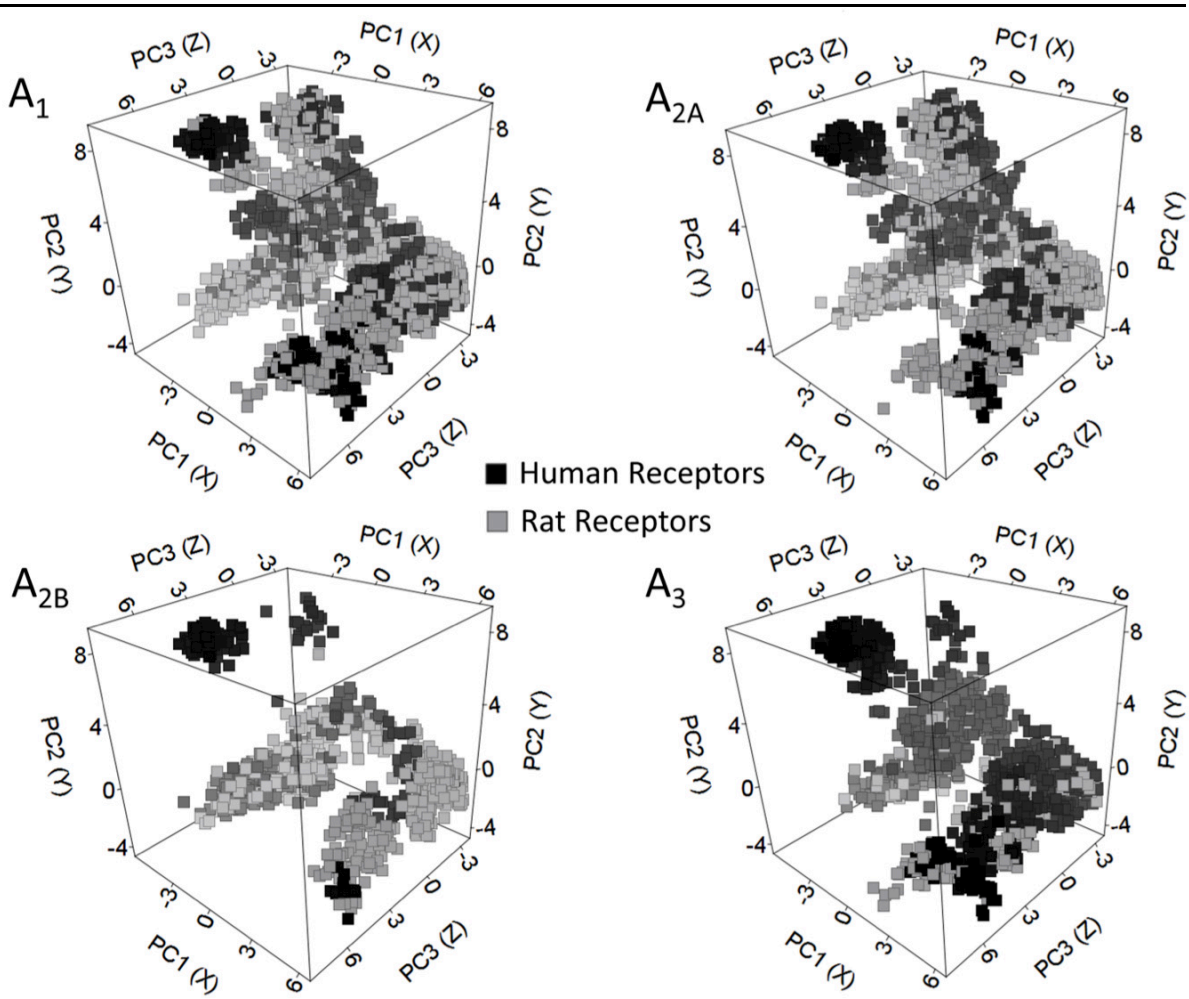


Figure 4.3: Principal component analysis of ligand chemical space. This PCA shows the large overlap in ligands that have been tested on ortholog pairs in the different species. The A_1 and A_{2A} receptors have the most densely populated chemical space, whereas A_{2B} has been explored the least. The space for the compounds tested on the A_3 receptors is dominated by compounds tested on the human ortholog. Note that the further along the x and z axes the point become lighter, black points fade to grey and grey points fade to white.

4.3.3 Target Descriptor. Firstly we identified the optimal selection method of the receptor binding site. (For a flow chart of the performed selections please see section **4.5**; here ECFP_4 fingerprints were employed as ligand descriptors, see section **4.5** section for further details.) In this part of the work we had a choice of 6 residue selection methods, two of which were structure based; the first one by selecting residues within a 5Å sphere around the co-crystallized ligand in PDB structure 3EML,[19] and the second one identical but using a 7 Å sphere. Furthermore, two selection methods were obtained utilizing TEA algorithms, selecting residues that were classified to be active in ligand recognition based on their evolutionary entropy.[20]

Here we used a conservative approach (TEA S), which selected a smaller number of residues, and a less restricted selection method (TEA L), which gave rise to a larger number of residues selected for model generation. Finally, we also evaluated two selection methods *combining* the a 5Å sphere with TEA S and one combining the 5Å sphere with TEA L. The best performing selection found was a combination of a 5 Å sphere around the co-crystallized ligand along with the small selection of TEA. This selection was named TEA S5 (Supporting **Figure S3** and Supporting **Table S3**).

During the optimization of our target descriptor by sampling different residue selection methods, we found that a larger selection is not always better. In fact, while the best performing binding site definition consisted of a combination of the two selection methods (crystal structure based and TEA based), it was in both cases the smallest residue selection within each method that performed optimally. Interestingly, we found that we needed to combine the crystal structure selection and TEA selection for optimal performance. In each individual selection method, both in the method based on the crystal structure alone and the method based solely on TEA, there was a pair of ortholog receptors that gave rise to an identical fingerprint. These were the A_{2A} receptors when selecting either a 5 Å or 7 Å sphere around the ligand in the crystal structure and the A_{2B} receptors when using the TEA based selection. However, since it was still possible to create predictive models in each individual case, it can be concluded that the activity space of these ortholog receptor pairs is highly similar (also see Supporting **Figure S1** and **S2**).

4.3.4 Ligand descriptor. Similar to the method used to identify the best descriptor binding site, we also identified the ligand descriptor giving rise to best modeling performance. Bender *et al.* in their analysis of descriptor space have shown that there is little difference between circular fingerprint performance and in this work similar results were obtained.[21] Our models identified the extended connectivity fingerprint using Sybyl atom typing (SCFP_4) to be the best-performing compound descriptor on this dataset (with an external validation RMSE of 0.70 log units and R_0^2 of 0.67) with three others close in performance. Those were FPFP_6 (RMSE 0.70 log units and R_0^2 0.68), EPFP_6 (RMSE 0.68 log units and R_0^2 0.69) and SPFP_6 (RMSE 0.69 log units and R_0^2 0.69) (see Supporting **Figure S4** and Supporting **Table S4**). Also in *predictive* power, *i.e.* the performance estimates in the cross validation compared to the external validation, the different fingerprints perform very similar but SCFP_4 better correlates to the external validation than in the others (see Supporting **Figure S4** and Supporting **Table S4**). We found this to be of high importance as we did not want to embark on a 'wet' experiment without having a fair estimate of model performance on unknown compounds.

4.3.5 Cross Validation. Finally, we sampled different cross-validation approaches by varying the amount of subdivisions in each cross-validation step. We observed that in the case of 5-fold cross validation the cross validation parameters are slightly worse compared with the external validation parameters, with a cross validated RMSE of 0.70 log units versus an RMSE for the external validation set of 0.68 log units. In addition the Q^2 is 0.69 in the cross validation and the R_0^2 is 0.71 in the external validation (Supporting **Figure S5**). When we increased the number of subdivisions, and hereby decreased the size of the fraction left out of the training during cross validation, this phenomenon was reversed. Hence the cross validated RMSE is slightly lower compared with RMSE in external validation (0.68 versus 0.70) and the Q^2 is slightly higher compared with R_0^2 (0.70 versus 0.69). This can indicate slight overtraining, as shown by Baumann,[22] which is the reason why we choose to implement 5-fold CV in the final model training procedure.

4.3.6 Final model training. The final model was trained on the full data set of eight receptors and 10,999 annotated data points. Given the preliminary results listed above, the model was built using SCFP_4 compound fingerprints and the TEA S5 residue selection. The training plot of the final model is shown in **Figure 4.4**, obtaining an R_0^2 of 0.95 and an RMSE of 0.26. The cross validated parameters, which constitutes a performance estimate, were a correlation coefficient (Q^2) of 0.73 and a prediction error (CV_RMSE) of 0.61 log units. This final model, created in Pipeline Pilot 8.5,[23] is provided in the Supporting Material. Furthermore, we also included in the Supporting Materials two tables showing the 25 substructures that have the largest positive (presence of these substructures leads to a higher pKi, Supporting **Table S5**) or largest negative effect on pKi (on average presence of these substructures leads to a lower pKi, Supporting **Table S6**). Finally, we have added the average effect on binding of the presence of the most occurring substructures. The top 100 most occurring substructures and their average effect on binding when present are given in Supporting **Table S7**.

The training plot of the final model, **Figure 4.4**, shows that especially compounds in the high pKi region (larger than 9) seem to be predicted more accurately. However, several compounds in this area, marked with a black circle, have been found to be underpredicted by a large margin. Upon identifying the outliers it was discovered these three points contained the same structure. Further literature studies showed that the outliers are all from Jacobson *et al.*[24] and that the original paper states that two of these compounds have been tested on the rat A_{2A} receptor, whereas ChEMBL 2 list them annotated to the human A_{2A} receptor. Moreover, the binding affinity values from this particular paper are much higher (pKi larger than 8.0) when compared with the affinity values these compounds and a large number of highly similar other compounds have on average in other papers (smaller than 7.0 and sometimes even smaller than 6.0). In Supporting **Table S8** are further details, i.e. the affinity of these particular compounds and a number of similar compounds in the training set. It would seem to be a reasonable explanation that these questionable values were experimental artifacts and a database annotation error. However, since this effect only occurs sporadically and only in the cases of these 8-cyclohexylcaffeine derivates we decided to keep the data points in the final model. It should be noted that our model was able to pick up these outlying experimental values. However; another consideration was that exhaustive checking of all 10,999 data points was practically infeasible.

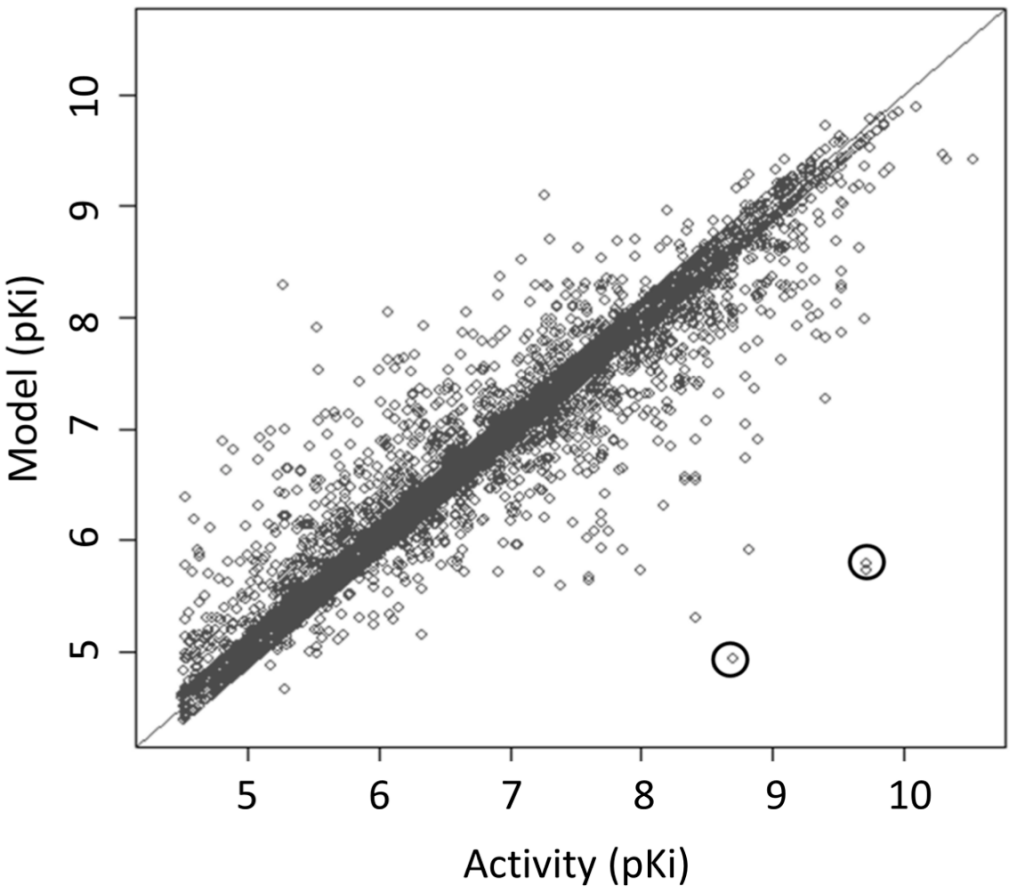


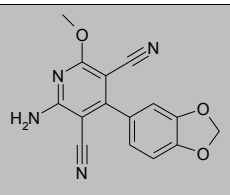
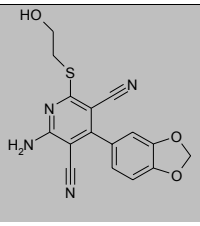
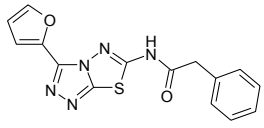
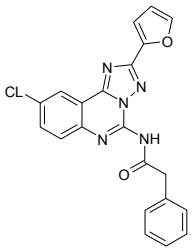
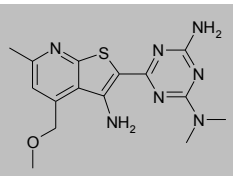
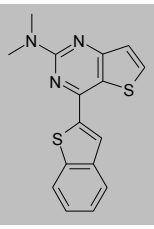
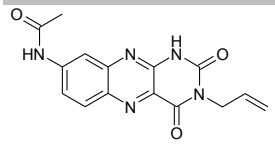
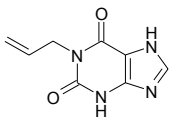
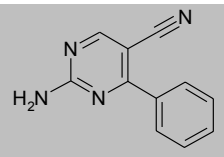
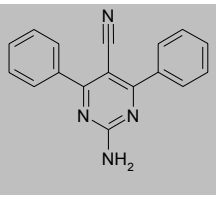
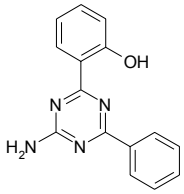
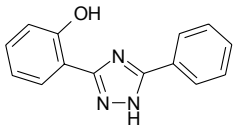
Figure 4.4: Cross validation plot of our final model correlating measured and predicted receptor affinities (pK_i values). The CV parameters were a Q^2 of 0.73 and a CV_RMSE of 0.61. The model fit had an R_0^2 of 0.95 and corresponding RMSE of 0.26. For an analysis of the outliers in the black circles see section **4.3.6**.

4.3.7 *In silico* model validation. Before applying our model in any virtual screening set-up, we performed several computational validation steps to ensure model predictivity and to prevent chance correlations from occurring. The learning curves (Supporting **Figure S6**) showed the maximal performance obtained was a prediction error of 0.62 log units (and corresponding R_0^2 of 0.71). In addition learning curves generated based on only the chemical space (Conventional structure-activity models rather than PCM models) showed that PCM is better able to model the ligand – target affinity than conventional single-target bioactivity models (Supporting **Figure S7** and Supporting **Figure S8**). The final model showed fair performance in external validation (Supporting **Figure S9**). It should be noted that our external validation consisted of compounds only tested on the human receptors. Interestingly the RMSE improved from 0.88 log units, when rat data was excluded from model training, to 0.82 log units, when these data were included (with the R_0^2 improving from 0.23 to 0.28).

Furthermore the model showed good performance in the decoy validation, since 33 of the 43 known actives were in the top 50 retrieved from 4,556 decoys (Supporting **Figure S10)**, with a runtime of 43 seconds. The highest predicted compound was LUF5957 with a predicted pKi of 9.02 (hA_1) and an experimentally determined pKi of 9.14 (hA_1). See Supporting **Table S9** for the structures of the four highest predicted decoys at rank 15, 21, 24 and 26. The 100-fold y-scrambled models plot shows a negative intersect with the Y-axis for both the R_0^2 and Q^2 regression lines as suggested to be characteristic for a predictive model by Eriksson *et al.* (Supporting **Figure S11**).[25]

The results from the different experiments show that the model appears to be statistically sound in nature. Several conclusions can be drawn from these results already. Firstly, it is difficult to train a model on public data gathered from a multitude of assays performed in different labs (also shown by Kramer *et al.*).[26] Secondly, our model is not based on chance correlations and has predictive power. Finally the pooling of data points from testing on rat receptors with data points from testing on human receptors has a positive effect on model performance. The RMSE improved from 0.88 to 0.82 upon inclusion of rat bioactivity data, likely by the inclusion of a much larger chemical space. Given the satisfying performance of our final model, we employed it in the next step to select novel potential adenosine receptor ligands from a chemical supplier, namely ChemDiv.

Table 4.1. Structures of the newly identified human adenosine receptor ligands.

Structure	Ki (μM, SEM and LE in parentheses) or % displacement at 10 μM				Most similar compound in training set	Similarity to training set (receptor)
	A_1	A_{2A}	A_{2B}	A_3		
 ***1***	**0.51 (±0.089, 0.39)**	**31 (±6.7, 0.28)**	32%	21%		0.65 (hA_1)
 ***2***	44%	**5.1 (±0.36, 0.32)**	-8%	3%		0.47 (rA_1)
 ***3***	35%	**1.6 (±0.31, 0.33)**	-5%	33%		0.30 (hA_1)
 ***4***	**3.2 (±0.13, 0.33)**	42%	14%	-3%		0.60 (rA_1)
 ***5***	**0.90 (±0.15, 0.55)**	**0.16 (±0.026, 0.62)**	-11%	13%		0.80 (hA_{2A})
 ***6***	**0.0072 (±0.0020, 0.56)**	**0.043 (±0.016, 0.51)**	**0.22 (±0.014, 0.46)**	**0.44 (±0.0073, 0.44)**		0.68 (hA_{2A})

Receptor affinity as determined in radioligand binding studies is shown as K_i value in μM or % displacement at 10 μM. Between parentheses the SEM in μM and the ligand efficiency (LE, see section **4.3.10**) in shown in kcal / mol per heavy atom. Also shown is the most similar compound in the training set (and the receptor it was annotated to) calculated as Tanimoto Similarity using the SCFP_4 fingerprint. Both entirely novel and atypical bioactive compounds have been identified (structures 3 and 4), as well as a fragment-like compound (structure 5) and a ligand with nanomolar activity (structure 6).

4.3.8 *In vitro* model validation. In our final 'wet' experimental validation we ordered 54 compounds that were indicated as active by our model on one or more of the adenosine receptors (see supporting SD file and Supporting **Table S10**). These 54 compounds were subsequently tested on all four human adenosine receptors (216 data points). Out of the total of 54 compounds tested six compounds were novel active compounds for the adenosine receptors (displacement larger 50% at a concentration of 10 µM; corresponding to a hit rate of 11%). Among the compounds were both selective ligands and highly active binders. For all six compounds active on either the human A_1 or human A_{2A} receptor in single-dose experiments full displacement curves were recorded, yielding K_i values. Furthermore, the pseudo Hill-coefficient was determined using variable slope regression in Graphpad Prism.[27] (The pseudo Hill-coefficients are listed in the supporting information along with all dose-response curves.)

Very diverse chemistry can be identified among the ligands found by our PCM model which are shown in **Table 4.1**. Two of the hits we found (compounds **1** and **2**) have a structure that resembles structures of known adenosine receptor ligands. However compound **3** and **4** have a structure that is not typical for compounds that are active on the adenosine receptors. Compound **5** shows a high affinity (0.90 µM on the human A_1 receptor and 0.30 µM on the human A_{2A} receptor), even though it is a very small fragment-like compound (MW 196). Finally, compound **6** even reached nanomolar affinities, even though no modifications or optimizations were performed on this compound. Note that the Tanimoto similarity to the training set based on the SCFP_4 fingerprint is as low as 0.30 in the case of compound **3** and reaching a maximum of 0.80 in the case of compound **5**. Furthermore, for two of the identified hits, the compound that is most similar in the training set has been annotated on the rat (A_1) receptor, further underlining the added value of the combination of human and rat orthologs. For additional details concerning the average and minimal similarity of the identified hits please see supporting **Table S11.** Shown in **Figure 4.5** are the curves used to determine the affinity of compound **6** on the human A_1 receptor. The full set of curves is contained in the supporting information.

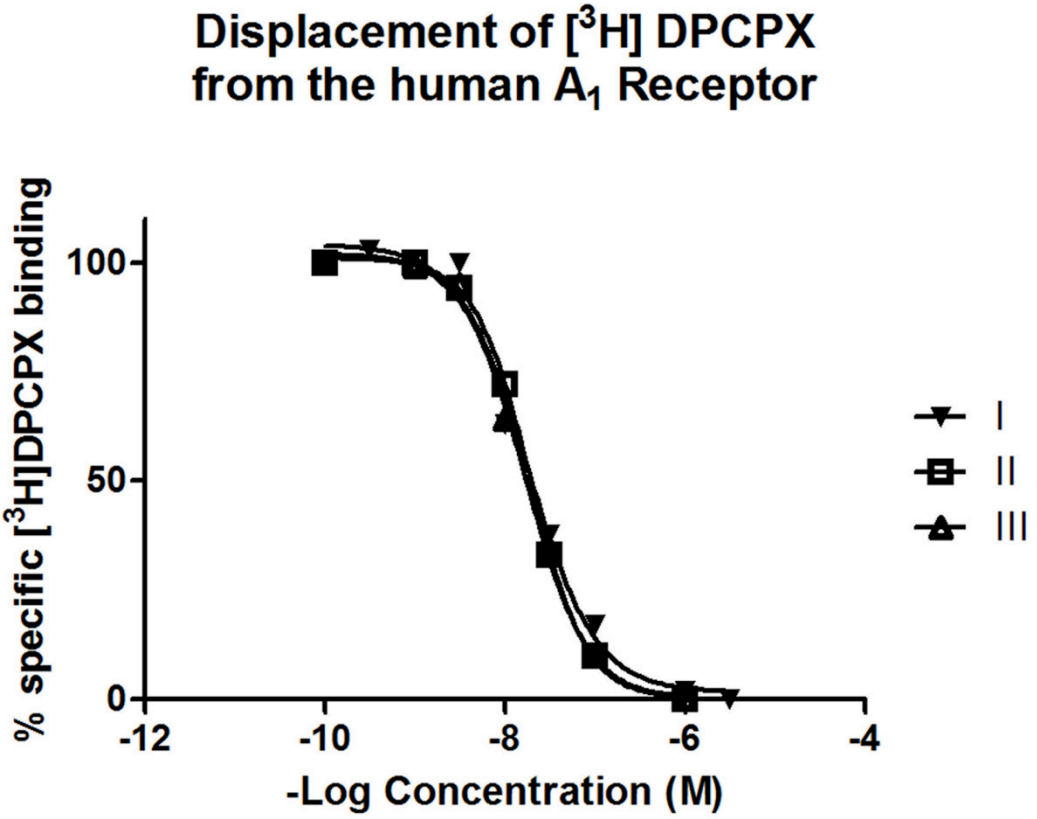


Figure 4.5: Typical dose response curve obtained during the in vitro model validation. Shown here are the dose response curves for compound **6** on the human A_1 receptor. The three curves performed in duplicate were obtained on different days. The pseudo Hill -coefficient was determined at -1.3 (± 0.1).

4.3.9 Implications on PCM performance. Since compounds **3** – **5** do not have a typical adenosine receptor template structure, we conclude that the PCM models obtained in this work are able to explore novel regions of bioactive chemical space. The ability of the model to find novel compounds is very likely the results of the larger chemical space covered in the training set in comparison with a conventional structure-activity model. (For a comparison of PCM and conventional structure-activity learning curves see Supporting **Figure S7** and Supporting **Figure S8**.) Together with the improved performance in the experimental validation, we show here the advantage PCM has due to its ability to characterize the full ligand – target interaction space.

However, while the PCM technique should in theory be able to predict bioactivity spectra, our experimental results indicate that our current model could not do so on the current data set. We were able to find active compounds and also selective compounds, but the compounds did not show selectivity as predicted by our model.

Furthermore, only one compound was found active on the human A_3 receptor, despite the fact that the model initially identified a much large number of compounds to have a pKi larger than 7.0 on the human A_3 receptor than on the other thee human receptors (see section **4.5.9** to **4.5.12** for details about compound selection). It is likely that this indicates that the model is not able to accurately model the bioactivity space for this receptor, in particular when we consider the large dissimilarity to the rat A_3 receptor.[1] The large dissimilarity combined with the low hit rate on the human A_3 receptor could indicate that the binding site definition is inaccurate. However it should be noted that this definition was based on only a single adenosine A_{2A} receptor crystal structure. As there are now more than a dozen GPCR crystal structures available, perhaps these can be used to better define the ligand binding site.

These two observations about the performance of the PCM model, the hit rate of 11% and the low performance for the human A_3 receptor, serve to illustrate that bioactivity models, like this model, are mainly a tool to *assist* in the process of medicinal chemistry. However this tool can be a very powerful tool as illustrated by the discovery of novel active compounds in the current work.

While this manuscript was completed, Langmead *et al.* published a structural virtual screening approach applied to the human adenosine A_{2A} receptor, identifying one out of 10 hits similar to compound **6,** subsequently optimized to be selective (for the structures see Supporting **Figure S12**).[28] It is interesting to see that we were able to identify a similar hit without the need for structural information.

4.3.10 Ligand Efficiency. Two of the identified novel ligands (**5** and **6**) showed a submicromolar affinity for (some of) the adenosine receptor subtypes. After calculating the ligand efficiency (LE),[29] we found that these two compounds both have a ligand efficiency higher than 0.5 kcal / mol per heavy atom on both the human A_1 and A_{2A} receptors. Furthermore, with the exception of compound **1** on the human A_{2A} receptor, all compounds have an LE higher than 0.30. Previously it has been shown that an LE of in the range of 0.30 – 0.40 constitutes a good value for lead optimization.[30-32] From the training set we also calculated the average LE (and standard deviation of this average) for ligands for each of the receptors, which was around 0.34 (supporting **Table S12**). These two compounds have a much higher LE, which renders them good starting points for the synthesis of a novel series now being pursued by our group.

4.3.11 PCM versus similarity searching. To place the performance of our PCM model in a broader context several similarity searching experiments were performed (supporting **Table S13**). To find 4 of the 6 hits, all compounds with a maximal (Tanimoto) similarity of 0.60 or higher should have been ordered, in ChemDiv this would have been 900 data points. However, the identification of all 6 hits would have required the purchase of all compounds with a maximal similarity of 0.30 or higher, a total of approximately 202,712 data points (on average approximately 50,000 compounds per human receptor). Moreover, the similarity searching was considerably slower than application of the PCM model. While the PCM model takes training time before it can be applied (3 hours on a Core i7 at 2.8 GHz with 16 GB of RAM), application afterwards is very quick, screening the full 791,162 compounds on all 4 receptors in 3 hours and 29 minutes (30 minutes for the filtered set; 100,910 compounds 15% of the total) using 6 threaded parallelization. Virtual screening using similarity searching on the same machine of all 791,162 compounds on the four human receptors took 71 hours and 29 minutes with a total time of 9 hours and 5 minutes for the filtered set (6 threaded calculation parallelization). The reason likely is that the PCM based approach requires a single calculation per data point (compound – receptor pair) whereas for the similarity searching, each compound requires the calculation of between 780 (hA_{2B}) and 1,661 (hA_3) Tanimoto similarities (indeed screening for the hA_{2B} receptor was considerably faster than the hA_3 receptor).

Likewise, we performed the decoy validation using a simple (Tanimoto) similarity based method (compounds ranked by maximal similarity to the training sets). Here we found that 40 of the known actives were in the top 50, with the highest predicted decoys at rank 35, 36, 38 and 43 (Supporting **Table S9**). However, while the PCM does not perform significantly better, the runtime for the similarity based approach was 6 minutes and 42 seconds (almost 10 times as long). In addition, similarity searching will not identify novel structures, which was the goal of this work.

These two similarity searching experiments demonstrate the added value of PCM as it displays a better performance and enrichment in prospective virtual screening combined with a significantly faster screening performance.

4.4 Conclusions

In this work we employed proteochemometric modeling (PCM) in order to identify novel human adenosine receptor ligands. By merging human and rat bioactivity data, we were able to identify six novel compounds that bind to members of the adenosine receptor family. One of these identified hits is very similar to a compound that was published recently (while we finalized this manuscript) using a structural rather than statistical approach. These novel ligands had an average Tanimoto similarity of 0.58 to the training set (ranging between 0.30 and 0.80). From the results we obtained we conclude that PCM is capable of capturing the full ligand – target space of a receptor subfamily rather than a single target. We showed that the addition of chemical and target information from orthologs can improve model quality when compared to creating a model based on a single species (prediction error decreased by 0.06 log units on this dataset). The ligand – targets spaces of human and rat adenosine receptors should not be regarded as separate entities and these spaces in fact overlap.

With the emergence and growth of large public databases such as ChEMBL,[15] PDB,[33] and Pubchem,[34] the PCM approach is likely to gain even further in momentum. In addition, the flexibility of this method may allow its application to other areas of drug discovery, such as receptor deorphanization or to different target families, such as the prediction of bioactivity profiles against kinases.

4.5 Experimental Section

4.5.1 Methods Overview. A flowchart of the modeling performed in this work is shown in **Figure 4.6**. The complete work can be divided in four major sections. Firstly, we created six different protein descriptors by varying the residue selection used to obtain them. Secondly, we created 16 different ligand fingerprints and also varied the maximal bond lengths in the substructures. Here we used a maximal diameter of either 4 or 6 bonds from a central atom. The third step consisted of finding the optimal combination of parameters for the training of the final model. These parameters included the different descriptors, method of cross validation and extended validation. The fourth and final step was the actual screening experiment where we combined both virtual screening and experimental validation. In the following section the individual procedures within each of these four blocks will be described starting with the descriptors for both the ligand and the receptors.

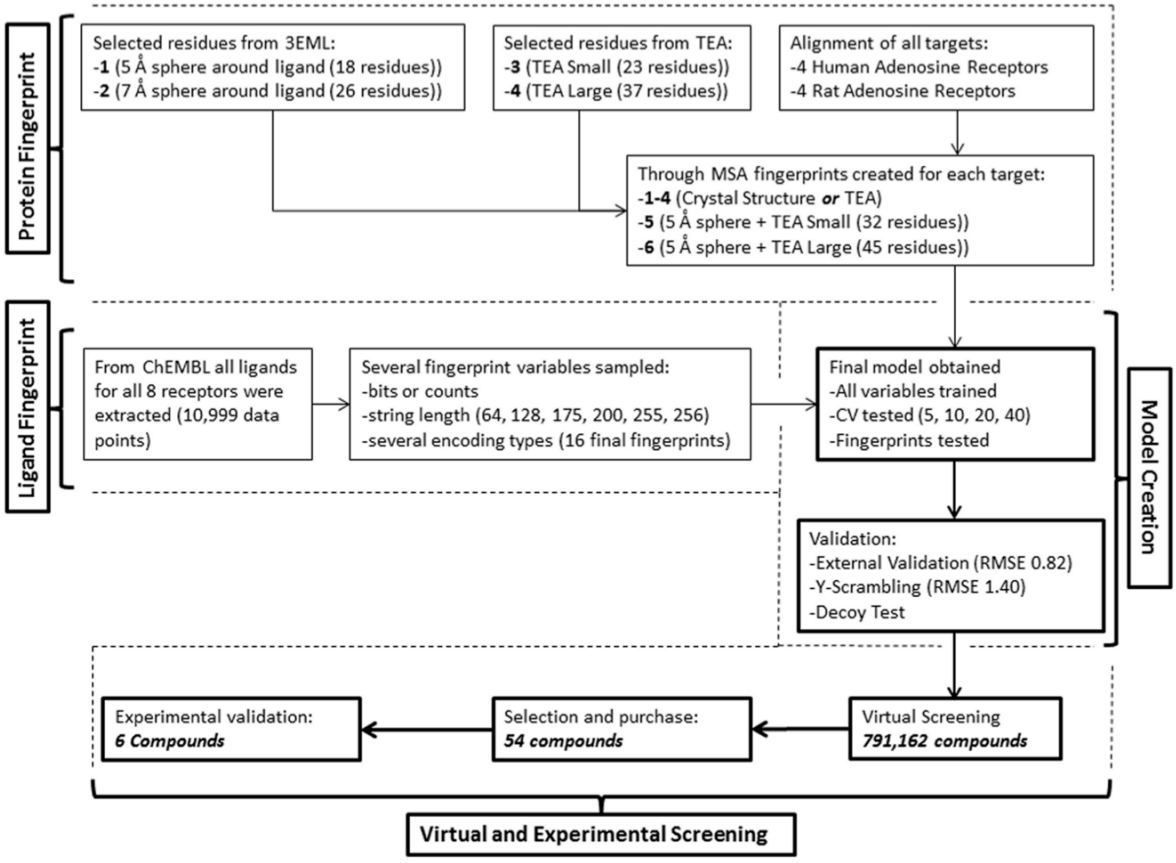


Figure 4.6: Flowchart of the work we performed.

4.5.2 Data set. The data set was obtained from ChEMBL 2.[15] From this database we selected all compounds that were tested on either human or rat adenosine receptors or both (Supporting Table **S2**). The selection was further narrowed to only include compounds for which a Ki value obtained from a radioligand binding assay was available. After selection the compounds were normalized and ionized at pH 7.4, they were assigned 2D coordinates and subsequently converted to fingerprints. All steps of this work were performed in Pipeline Pilot Student Edition version 6.1.5.[35]

The receptor sequences were obtained from Uniprot and aligned using ClustalW (Slow alignment, Gap Open 4, Gap Extend 4, available as supporting information).[36] This alignment was used to convert residues selected from the crystal structure to their ortholog and paralog counterparts. The residues selected by the TEA approach are provided in Ballesteros-Weinstein numbers and could be used directly.[37] After selection of the residues they were converted to a feature based protein fingerprint based on their single letter amino acid codes as we have done in previous work.[10]

4.5.3 Descriptor Benchmarking Approach. Before we could train our final predictive model, we sampled a multitude of parameters. We collected 6 selection methods to define our binding site residues (Supporting **Figure S3** and **Table S3**), 16 different types of circular fingerprints (Supporting **Figure S4** and **Table S4**), and 4 different folds of cross validation (CV) (Supporting **Figure S5**). From these options we wanted to select the optimal combination of variables. To identify the best combination, all models were built on 70% of the dataset (7,749 data points) and validated on the remaining 30% (3,250 data points). From the learning curves we already knew that this split was the optimal partition (Supporting **Figure S6**).

4.5.4 Protein descriptors. Sequences were encoded based on the binding site sequence in which each amino acid was represented as a single unique feature as was done in previous work.[10] However, these residues were selected in seven different ways and each selection was tested to find the best option to be used in the final model (Supporting **Figure S3**). The first two selection methods (1 and 2 in **Figure 4.6** were based on the crystal structure of the adenosine A_{2A} receptor bound to ZM241385 (PDB code 3EML).[19] Herein all residues were selected having any atom within either a 5 Å or a 7 Å sphere around the co-crystalized ligand. The third and fourth selection methods (3 and 4 in **Figure 4.6**) were based on a bioinformatics approach known as Two Entropy Analysis (TEA).[20] This method relies on quantifying the degree to which trans-membrane (TM) residues are conserved among Class-A GPCRs. Both the degree of conservation among GPCR subfamilies and the degree of conservation among the whole family are calculated. This calculation then serves as a basis to differentiate the function residues perform in individual GPCRs based on the difference between these degrees of conservation. Here we used a conservative approach (TEA S), which was small, and a less restricted selection method (TEA L), which was larger and included some of the residues from the 'mixed region' mentioned in the original publication.[20] The fifth method (TEA S5, 5 in **Figure 4.6**) was based on a combination of 1 one and 3, the sixth method (TEA L5, 6 in **Figure 4.6**) was based on a combination of method 1 and 4. Finally the seventh method consisted of simply using al TM residues. During this optimization, the ligand descriptor ECFP_4 was used (see **4.5.5** for further details).

The features describing the binding site were obtained by hashing an array of 58 physicochemical properties obtained from the AAindex database;[38] the indices employed can be found in Supporting **Table S14**. Finally, protein fingerprints were converted to an array of 175 features (Supporting **Figure S13**) which were then used in the modeling using Pipeline Pilot version 8.5.[23]

4.5.5 Compound descriptors. All descriptors were calculated in the academic version of PipelinePilot 6.1.5.[35] In the final model, ligands were described by Scitegic circular fingerprints (SCFP_4 type),[39, 40] which have previously been shown to capture a large amount of information with respect to compound bioactivity.[21] SCFP_4 descriptors provide individual substructures and treat these as a feature of a compound. We found them to perform the best and most consistent (Supporting **Figure S4**). These substructures have a maximal diameter of 4 bonds from a central atom. Finally ligand fingerprints were converted to an array of 175 features which were then used in the modeling (Supporting **Figure S13**).

4.5.6 Machine learning. Models were constructed in the academic version of Pipeline Pilot 6.1.5 using the R-statistics package. Support vector machines (SVM) as coded in the e1071 package were used for model creation.[41] Parameters gamma and cost were tuned over an exponential range and epsilon was set at 0.1. The optimal model was determined using cross validation before proceeding to experimental prospective validation of the model. The parameters used for validation were R_0^2, R^2, and RMSE.[42, 43]

4.5.7 *In silico* validation. We performed four different *in silico* validation experiments. Firstly a learning curve was generated, to spot possible discontinuous randomized splits data points, prevent overtraining and obtain an estimate of maximal performance that can be obtained on this dataset (Supporting **Figure S6**). Secondly, the obtained final model was subjected to external validation, applying the model to previously unseen compounds not part of our training set (**Supporting Figure S9**). Thirdly, the model was applied to a decoy set validation, to check performance in identifying unseen known actives from decoys (Supporting **Figure S10**).

This decoy set consisted of random selection of compounds from the ZINC database selected to resemble adenosine receptor ligands based on their physicochemical properties. These properties included molecular weight, number of hydrogen bond donors / acceptors, number of aromatic rings, calculated AlogP, average bond length, number of atoms, number of rotatable bonds, and formal charge (Supporting **Figure S14**). Finally y-scrambling was performed to estimate the possibility of chance correlations (Supporting **Figure S11**).

4.5.8 Virtual screening. Subsequent to our model validation we performed a virtual screening. We screened all compounds available in the ChemDiv database obtained *via* ZINC (accessed December 3rd 2009 consisting of 791,162 compounds) without any form of pre-filtering as we wanted a fair estimate of true model performance.[44] The main advantage of our statistical method is its throughput; typically one can screen the full ChemDiv database within 4 hours on a desktop machine (core i7 at 2.8 GHz with 16 GB RAM). In this case we screened 791,162 compounds on the four human Adenosine receptors (3,164,648 data points).

4.5.9 Selection Filters. On the resulting model output, several filters were applied: molecular solubility larger than -4 (solubility expressed as logS with S in mol/L),[45] AlogP between -0.4 and 5.6.[46] This reduced the number of data points from 3,164,648 to 403,640. The filtering could also very well have been applied before screening, but we wanted to see if our technique was capable of screening such a large number of compounds in a reasonable time. While more than 400,000 data points still represented a significant number the next step was binning and diversity clustering based on the chemistry of the compounds.

4.5.10 Binning. Subsequent to the classification ranking, compounds were binned in the following classes: predicted pKi between 5.0 and 6.0 (Bin 1; 180,419 data points), predicted pKi between 6.0 and 7.0 (Bin 2; 19,314 data points) and predicted pKi larger than 7.0 (Bin 3; 2,875 data points). The remainder was predicted to have a pKi smaller than 5.0 (201,032 data points) and were discarded together with those compounds not meeting the earlier physicochemical filters.

4.5.11 Clustering. Clustering was subsequently performed using the pipeline pilot component 'Cluster Molecules' on the individual bins with the aim of creating subsets containing different chemistry (Bin 1: 10 clusters for each receptor, Bin 2: 9 clusters for each receptor and Bin 3: 6 clusters for each receptor). The descriptor used was identical to the one we used to train the final models and the similarity coefficient used as the Tanimoto coefficient. For details about the obtained clusters see Supporting **Table S15** to **Table S18**. Bin 3 was to serve as a pool to select compounds for experimental validation. Note that Bin 3 is much larger in the case of the human adenosine A_3 receptor, this could indicate that the model is better able to find high affinity compounds for the human A_3 receptor. However, it is more likely that this indicates that the model is not able to accurately model the bioactivity space for this receptor, in particular when we consider the previously low hit rate in other studies on that receptor and the large dissimilarity to the rat A_3 receptor.

4.5.12 Final compound selection. Finally, compounds to be ordered were selected manually from Bin 3 for each receptor (predicted pKi larger than 7) with a focus on the selection of novel chemotypes. In total 54 compounds were selected. (Supporting **Table S10**, and supporting SD file).

These compounds were selected from different clusters (these clusters are indicated in bold in Supporting **Tables S15**-**S18** and are also present in the supporting SD file). The ignored clusters represented clusters that contained the remaining compounds (junk clusters) in the case of the human A_1, human A_{2A} and human A_3 receptor. In the case of human A_{2B} receptor the ignored cluster contained a number of compounds that were chemically very similar to compounds already selected for the hA_1 and hA_{2A} receptor (and were hence already going to be tested in the hA_{2B} receptor).

The compounds selected included both compounds that were predicted to be active on multiple receptors (like compound 5, predicted to be active on all 4 human receptors) and compounds predicted to be selective (like compound 3 predicted to be active on hA_{2A} and hA_{2B}). These compounds were subsequently ordered and tested *in vitro* on all receptors (216 data points), one compound (55 in Supporting **Table S10**) could not be ordered as it was unavailable from the supplier. 1H NMR and MS data are included in the Supporting Information for the found hits.

4.5.13 Ligand efficiency. Ligand efficiency (LE),[29, 31, 47] expressed in kcal / mol per heavy atom, was calculated according to eq. 1.

$$LE = \Delta G / N_{Non\text{-}hydrogen\ atoms} \quad (1)$$

To obtain ΔG we used eq. 2. ΔG was converted to kcal / mol.

$$\Delta G = -RT.\ln K_i \quad (2)$$

4.5.14 Similarity searching. All similarity searching experiments were performed in Pipeline Pilot version 8.5.[23] The Pipeline pilot similarity searching component was used and the search was done using SCFP_4 fingerprints. The component was optimized for speed rather than memory use and screening was done in parallel using 6 threads on a core i7 machine. For each receptor subtype, the subset of the training set regarding that subtype was used as reference compounds. For example, to identify similar compounds for the human A_1 receptor, we used the human A_1 receptor annotated compounds from ChEMBL 2.

4.5.15 Binding Studies. [3H]DPCPX and [3H]ZM241385 (4-(2-[7-amino-2-(2-furyl)[1,2,4]triazolo[2,3-α][1,3,5]triazin-5-ylamino]ethyl)phenol) were purchased from ARC Inc. St Louis, USA. [3H]PSB603 and [3H]PSB11 were kind gifts from Prof C.E. Müller (Bonn, Germany). Chinese Hamster Ovary (CHO) cells expressing the human adenosine A_1 receptor were provided by Dr. Andrea Townsend-Nicholson, University College of London, UK. Human embryonic kidney (HEK) 293 cells stably expressing the human adenosine A_{2A} and human A_3 receptor were gifts from Dr. Wang (Biogen) and Dr. K.-N. Klotz (University of Wűrzburg, Germany), respectively. CHO cells expressing the human A_{2B} receptor were provided by Dr. Steve Rees (GlaxoSmithKline, UK). Dose response curves for the found hits are included in the Supporting Information.

4.5.16 Human adenosine A_1 Receptor. Affinity at the A_1 receptor was determined on membranes from CHO cells expressing the human receptors, using [3H]DPCPX as the radioligand. Membranes containing 5 µg of protein were incubated in a total volume of 100 µL of 50 mM Tris•HCl (pH 7.4) and [3H]DPCPX (final concentration 1.6 nM) for 1 h at 25 °C in a shaking water bath. Nonspecific binding was determined in the presence of 100 µM CPA. The incubation was terminated by filtration over pre-wetted Whatman GF/B filters under reduced pressure with a Brandel harvester. Filters were washed three times with ice-cold buffer and placed in scintillation vials. Emulsifier Safe (3.5 mL) was added, and after 2 h radioactivity was counted in a TriCarb 2900TR liquid scintillation counter.

4.5.17 Human adenosine A_{2A} Receptor. At the A_{2A} receptor, affinity was determined on membranes from HEK 293 cells stably expressing this human receptor, using [3H]ZM241385 as the radioligand. Membranes containing 40 µg of protein were incubated in a total volume of 100 µL of 50 mM Tris•HCl (pH 7.4) and [3H]ZM241385 (final concentration 1.7 nM) for 2 h at 25 °C in a shaking water bath. Nonspecific binding was determined in the presence of 100 µM CGS21680. The incubation was terminated by filtration over pre-wetted Whatman GF/B filters under reduced pressure with a Brandel harvester. Filters were washed three times with ice-cold buffer and placed in scintillation vials. Emulsifier Safe (3.5 mL) was added, and after 2 h radioactivity was counted in a TriCarb 2900TR liquid scintillation counter.

4.5.18 Human adenosine A_{2B} Receptor. At the A_{2B} receptor, radioligand displacement was determined on membranes from CHO cells stably transfected with human A_{2B} receptor, using [3H]PBS603 as the radioligand. Membranes containing 15 µg of protein were incubated in a total volume of 100 µL of 50 mM Tris•HCl (pH 7.4), 1U/mL ADA, 0.1 w/v % CHAPS (pH 8.2 at 5 °C), and [3H]PBS603 (final concentration 1.0 nM) for 2 h at 25 °C in a shaking water bath. Nonspecific binding was determined in the presence of 100 µM NECA. The incubation was terminated by filtration over pre-wetted Whatman GF/C filters under reduced pressure with a Brandel harvester. Filters were washed three times with ice-cold 50 mM Tris•HCl, pH7.4 + 0.1% BSA buffer and placed in scintillation vials. Emulsifier Safe (3.5 mL) was added, and after 5 h radioactivity was counted in a TriCarb 2900TR liquid scintillation counter.

4.5.19 Human adenosine A_3 Receptor. The affinity at the A_3 receptor was measured on membranes from HEK 293 cells stably expressing the human A_3 receptor, using [3H]PSB11 as the radioligand. Membranes containing 25 µg of protein were incubated in a total volume of 100 µL of 50 mM Tris•HCl, 10 mM MgCl2, 1 mM EDTA, 0.01% CHAPS (pH 7.4), and [3H]PSB11 (final concentration 4 nM) for 1 h at 37 °C in a shaking water bath. Nonspecific binding was determined in the presence of 100 µM R-PIA. The incubation was terminated by filtration over pre-wetted Whatman GF/B filters under reduced pressure with a Brandel harvester. Filters were washed three times with ice-cold buffer and placed in scintillation vials. Radioactivity was counted in a Wallac 1470 Wizard gamma counter.

4.5.20 Data Analysis. K_i values were calculated using a nonlinear regression curve-fitting program (GraphPad Prism 5.0).[27] K_i values of radioligands were 1.6, 1.7, 0.41, and 4.9 nM for [^{3}H]DPCPX, [^{3}H]ZM241385, [^{3}H]PSB603, and [^{3}H]PSB11, respectively.

4.6 Supporting Information

Additional tables (Supporting **Tables S1 – S18**), figures (**Figures S1 – S14**), compound purity information (1H NMR, LC-MS and HR MS spectra), radioligand displacement curves for the found hits, the final model, an SD file with the ordered compounds, the multiple sequence alignment used to create the protein descriptor and a protocol to run the model are available online. These materials are available online at www.gjpvanwesten.nl.

4.7 Acknowledgments

GvW would like to thank Tibotec BVBA for funding. The authors would like to thank Jacobus van Veldhoven and Maris Vilums for help in analytical chemistry and Laura Heitman for the helpful discussions.

4.8 References

1. B.B. Fredholm, A.P. IJzerman, K.A. Jacobson*, et al.*; *International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and Classification of Adenosine Receptors—An Update.* Pharmacol. Rev.; *2011*. **63** (1): 1-34.
2. R.M. Hyde and D.J. Livingstone; *Perspectives in QSAR: Computer chemistry and pattern recognition.* J. Comput.-Aided Mol. Des.; *1988*. **2**: 145-155.
3. K. Roy; *QSAR of Adenosine Receptor Antagonists II.* QSAR Comb. Sci.; *2003*. **22** (6): 614-621.
4. A. Bender and R.C. Glen; *Molecular similarity: a key technique in molecular informatics.* Org. Biomol. Chem.; *2004*. **2**: 3204-3218.
5. A. Kontijevskis, R. Petrovska, I. Mutule*, et al.*; *Proteochemometric analysis of small cyclic peptides' interaction with wild-type and chimeric melanocortin receptors.* Proteins: Struct., Funct., Bioinf.; *2007*. **69** (1): 83-96.
6. M. Lapinsh, P. Prusis, A. Gutcaits*, et al.*; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
7. G.J.P. Van Westen, J.K. Wegner, A.P. IJzerman*, et al.*; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.
8. J. Wikberg, M. Lapinsh, and P. Prusis; *Proteochemometrics: A tool for modelling the molecular interaction space;* in *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*; H. Kubinyi and G. Müller; Editors. *2004*. p. 289 - 309.
9. M. Lapinsh, P. Prusis, S. Uhlen*, et al.*; *Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions.* Bioinformatics; *2005*. **21** (23): 4289-4296.
10. G.J.P. Van Westen, J.K. Wegner, P. Geluykens*, et al.*; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.
11. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.
12. B.B. Fredholm, A.P. IJzerman, K.A. Jacobson*, et al.*; *International Union of Pharmacology. XXV. Nomenclature and Classification of Adenosine Receptors.* Pharmacol. Rev.; *2001*. **53** (4): 527-552.

13. R. Guha and J.H. VanDrie; *Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs.* J. Chem. Inf. Model.; *2008*. **48** (3): 646-658.

14. M.T. Sisay, L. Peltason, and J.r. Bajorath; *Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure-Activity Relationships in Analog Series.* J. Chem. Inf. Model.; *2009*. **49** (10): 2179-2189.

15. A. Gaulton, L.J. Bellis, A.P. Bento*, et al.*; *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res.; *2011*. **40**: D1100 - D1107.

16. F.A. Kruger and J.P. Overington; *Global Analysis of Small Molecule Binding to Related Protein Targets.* PLoS Comput. Biol.; *2012*. **8** (1): e1002333.

17. J.E.S. Wikberg, F. Mutulis, I. Mutule*, et al.*; *Melanocortin receptors: Ligands and proteochemometrics modeling;* in *Melanocortin System*; D. Braaten; Editor *2003*: New York. p. 21-26.

18. M. Lapinsh, P. Prusis, T. Lundstedt*, et al.*; *Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.* Mol. Pharmacol.; *2002*. **61** (6): 1465-1475.

19. V.P. Jaakola, M.T. Griffith, M.A. Hanson*, et al.*; *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science; *2008*. **322** (5905): 1211-1217.

20. K. Ye, E.W.M. Lameijer, M.W. Beukers*, et al.*; *A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **63** (4): 1018-1030.

21. A. Bender, J.L. Jenkins, J. Scheiber*, et al.*; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.

22. K. Baumann; *Cross-validation is dead. Long live crossvalidation! Model validation based on resampling.* J. Cheminf.; *2010*. **2** (Suppl 1): O5.

23. Accelrys Software Inc *Pipeline Pilot Professional Edition* Scitegic Version 8.5

24. K.A. Jacobson, P.J.M. Van Galen, and M. Williams; *Adenosine receptors: pharmacology, structure-activity relationships, and therapeutic potential.* J. Med. Chem.; *1992*. **35** (3): 407-422.

25. L. Eriksson, J. Jaworska, A.P. Worth*, et al.*; *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environ. Health Perspect.; *2003*. **111** (10): 1361-1375.

26. J.R. Lane, C. Klein Herenbrink, G.J.P. Van Westen*, et al.*; *A Novel Nonribose Agonist, LUF5834, Engages Residues That Are Distinct from Those of Adenosine-Like Ligands to Activate the Adenosine A2a Receptor.* Mol. Pharmacol.; *2012*. **81** (3): 475-487.

27. GraphPad Software Inc *GraphPad Prism* 5.0

28. C.J. Langmead, S.P. Andrews, M. Congreve*, et al.*; *Identification of Novel Adenosine A2A Receptor Antagonists by Virtual Screening.* J. Med. Chem.; *2012*. **55**: 1904 - 1909.

29. P.R. Andrews, D.J. Craik, and J.L. Martin; *Functional group contributions to drug-receptor interactions.* J. Med. Chem.; *1984*. **27** (12): 1648-1657.

30. A.L. Hopkins, C.R. Groom, and A. Alex; *Ligand efficiency: a useful metric for lead selection.* Drug Discov. Today; *2004*. **9** (10): 430-431.

31. D. Tanaka, Y. Tsuda, T. Shiyama*, et al.*; *A Practical Use of Ligand Efficiency Indices Out of the Fragment-Based Approach: Ligand Efficiency-Guided Lead Identification of Soluble Epoxide Hydrolase Inhibitors.* J. Med. Chem.; *2010*. **54** (3): 851-857.

32. C. Abad-Zapatero and J.T. Metz; *Ligand efficiency indices as guideposts for drug discovery.* Drug Discov. Today; *2005*. **10** (7): 464-469.

33. H.M. Berman, J. Westbrook, Z. Feng*, et al.*; *The Protein Data Bank* Nucleic Acids Res.; *2000*. **28**: 235-242.

34. E.E. Bolton, Y. Wang, P.A. Thiessen*, et al.*; *PubChem: Integrated Platform of Small Molecules and Biological Activities;* in *Annual Reports in Computational Chemistry*; A.W. Ralph and C.S. David; Editors. *2008*; Elsevier. p. 217-241.

35. Accelrys Software Inc *Pipeline Pilot Student Edition* Scitegic Version 6.1.5

36. E. Jain, A. Bairoch, S. Duvaud*, et al.*; *Infrastructure for the life sciences: design and implementation of the UniProt website.* BMC Bioinformatics; *2009*. **10** (1): 136-155.

37. J.A. Ballesteros and H. Weinstein; *Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors;* in *Methods in Neurosciences*; C.S. Stuart; Editor *1995*; Academic Press. p. 366-428.

38. S. Kawashima, H. Ogata, and M. Kanehisa; *AAindex: Amino Acid Index Database.* Nucleic Acids Res.; *1999*. **27** (1): 368-369.

39. R.C. Glen, A. Bender, C.H. Arnby*, et al.*; *Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME.* IDrugs; *2006*. **9** (3): 199 - 204.

40. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.

41. E. Dimitriadou, K. Hornik, F. Leisch*, et al. Misc Functions of the Department of Statistics (e1071)* TU Wien 2006 1.5-15

42. A. Tropsha, P. Gramatica, and Vijay K. Gombar; *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.* QSAR Comb. Sci.; *2003*. **22** (1): 69-77.

43. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

44. J.J. Irwin and B.K. Shoichet; *ZINC − A Free Database of Commercially Available Compounds for Virtual Screening.* J. Chem. Inf. Model.; *2005*. **45** (1): 177-182.

45. I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva*, et al.*; *Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices.* J. Chem. Inf. Comput. Sci.; *2001*. **41** (6): 1488-1493.

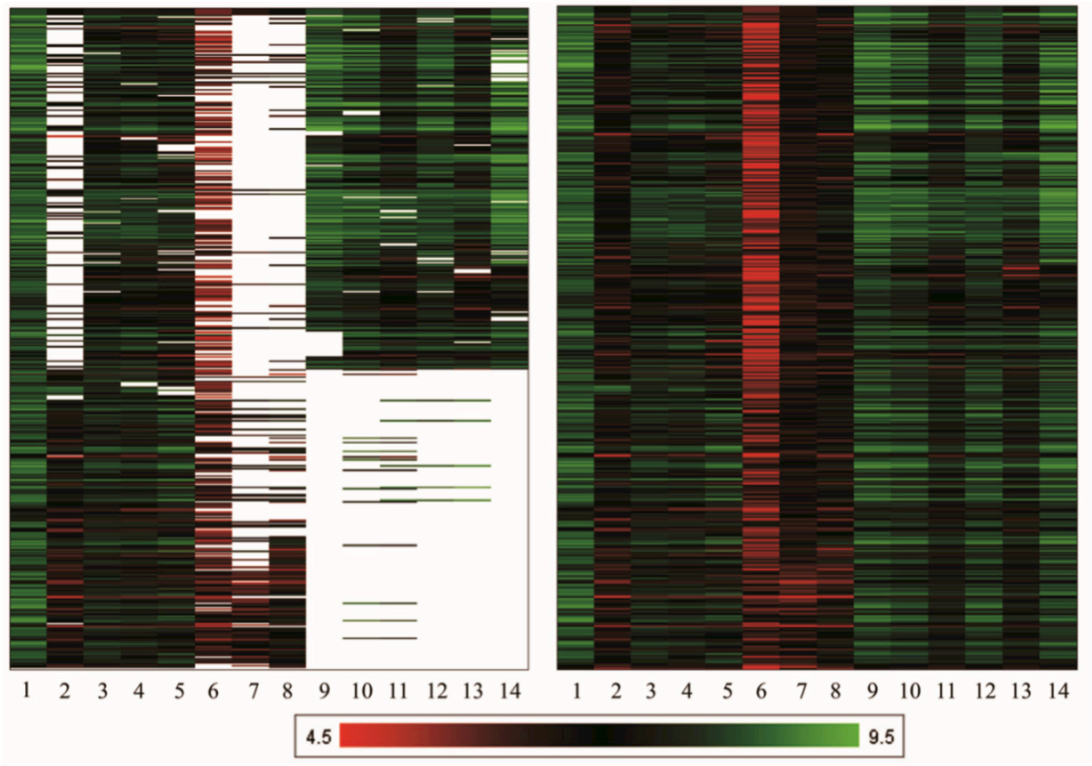
46. A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski; *Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods.* J. Phys. Chem.; *1998*. **102** (21): 3762-3772.

47. I.D. Kuntz, K. Chen, K.A. Sharp*, et al.*; *The maximal affinity of ligands.* Proc. Natl. Acad. Sci. U. S. A.; *1999*. **96** (18): 9997-10002.

Chapter 5

Which Compound to Select in Lead Optimization?

Prospectively Validated Proteochemometric Models Guide Preclinical Development



G.J.P. Van Westen, J.K. Wegner, P. Geluykens, L. Kwanten, I. Vereycken, A. Peeters, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; PLoS One; 2011. ***6*** *(11): e27518.*

Contents

5.1 Abstract

In quite a few diseases drug resistance due to target variability poses a serious problem in pharmacotherapy. This is certainly true for HIV, and hence, it is often unknown which drug is best to use or to develop against an individual HIV strain. In this work we applied 'proteochemometric' modeling of HIV Non-Nucleoside Reverse Transcriptase (NNRTI) inhibitors to support preclinical development by predicting compound performance on multiple mutants in the lead selection stage. Proteochemometric models are based on both small molecule and target properties and can thus capture multi-target activity relationships simultaneously, the targets in this case being a set of 14 HIV Reverse Transcriptase (RT) mutants. We validated our model by experimentally confirming model predictions for 317 untested compound – mutant pairs, with a prediction error comparable with assay variability (RMSE 0.62). Furthermore, dependent on the similarity of a new mutant to the training set, we could predict with high accuracy which compound will be most effective on a sequence with a previously unknown genotype. Hence, our models allow the evaluation of compound performance on untested sequences and the selection of the most promising leads for further preclinical research. The modeling concept is likely to be applicable also to other target families with genetic variability like other viruses or bacteria, or with similar orthologs like GPCRs.

5.2 Introduction

5.2.1 Genetic Information is readily available. Over the last decade extensive sequencing efforts have unraveled the human genome and provide an insight into the extent of human genetic variation.[1, 2] On the one hand this provides possible new drug targets that can lead to new drugs;[3-5] on the other hand it shows clearly that natural genetic variation needs to be addressed by some form of personalized medicine that works in a particular patient.[6] An exhaustive, individual "pharmacogenomics" approach for a patient, taking the full genetic make-up of a human into account, is unfortunately not feasible in the short term. This is due to the cost of sequencing but even more so to insufficient understanding of biological processes in humans.[7] However, what is already feasible today for every patient is the full sequencing of pathogens such as bacteria and viruses, as these contain a significantly smaller genome with relatively established locations and functions of drug targets. It is now possible through 'Deep Sequencing' technologies, to identify dominant and subdominant viral strains present in an individual patient, paving the way for the development of HIV inhibitors with an optimal potency profile made to target all relevant HIV variants.[8, 9]

What is required for the development of an optimal preclinical candidate on the other hand is a knowledge base of the effect mutations have on the binding of current inhibitors. When this information is available it can be used to create a model that allows the user to extrapolate between target sequence variants and predict binding affinities of preclinical compounds on previously untested viral target sequences. While similar models have been trained on this data for clinical drugs, these models have in common that they solely are trained on recognizing patterns of the presence and absence of mutations, thus only considering target information.[10-15] They do not take into account structural information of the compound – target interaction; hence they are not able to rationalize why an inhibitor is active on one sequence but not on another. As a result, the application to the discovery of preclinical candidates is rather limited.

5.2.2 How to Choose the Right Drug for a Genotype? In the current work we present one approach to remedy the situation, by making use of the large amount of structural data available on the binding of HIV Reverse Transcriptase (RT) inhibitors to their targets. We will show using prospective experimental validation on hundreds of data points that we can indeed predict which compound is preferable with regard to activity against particular mutants, compared to other compounds.

In particular, our aim was to predict activity of compounds on previously untested genetic variants of the virus. Given our in-depth understanding of the structural differences between viral enzyme sequences we can incorporate this knowledge to arrive at much improved extrapolation abilities, which enables the design of new inhibitors with improved broad activity profiles.

5.2.3 Extrapolating in Target Space. When learning from bioactivity data, and attempting to make predictions for novel chemical structures, statistical and machine learning techniques have a proven ability to 'make sense' of large data sets under certain conditions (such as interpretable variables used in the model) and to relate chemical structure to activity against a protein target. Bioactivity models are generally based on the 'Molecular Similarity Principle' stating that similar compounds (individual compounds or with respect to the distribution of chemistry in a given data set) possess similar properties, such as in this case similar bioactivit.[16-18] Yet conventional bioactivity models possess a severe limitation when considering sets of targets, which may be members of a target family such as kinases or G protein-coupled receptors (GPCRs), or as in the case presented here, sequences of viral enzymes. Those models take into account multiple molecules active on a single protein target, yet they completely neglect our extensive knowledge on the similarities of targets to each other. Hence, conventionally a single bioactivity model is generated for every target – neglecting that not only similar compounds show similar bioactivity, but reversely also that similar targets bind similar compounds. In addition, this concept is crucial for the chemogenomics paradigm that has been receiving lots of attention recently.[19, 20] In practice this means that even if a particular ligand-protein target data point is unknown, we can often extrapolate from neighboring activities in both ligand and target space, and not only in ligand space as has previously been done.

In this work we employed a modeling technique called 'proteochemometric modeling' (PCM) to model bioactivity data on a set of 14 enzyme sequences. This technique was introduced by Lapinsh, Wikberg et al.,[21, 22] and similar approaches have since appeared.[23] These techniques have been recently reviewed by the authors.[24] However, no large prospective studies have been presented until today. As PCM uses both ligand and target information, the hypothesis of our work was that PCM would be able to extrapolate the activity of compounds encountered in the training set to novel target sequences.

The extension of bioactivity modeling and its impact on preclinical drug research, plus the extensive prospective experimental validation performed, is the main contribution of the current work to the bioactivity modeling field. In addition, applications of PCM to novel target families including, but not limited to: Class A, B and C GPCRs; Kinases; Voltage-gated ion channels and others, can also be covered by this concept. The flexibility of the method is of particular interest when taking current multi-target drug paradigms into account.[25, 26]

5.3 Methods

5.3.1 Data set used to build the models. The data set employed comprises 451 compounds and 14 HIV RT sequences and hence 6,314 possible compound – target combinations. The set was generously provided by Tibotec BVBA. For a total of 4,024 of these combinations an activity value in the form of a pEC_{50} value was available for training the bioactivity model. Compounds with a pEC_{50} value on a certain sequence closer than 0.3 log units to the toxic concentration for that compound (expressed as pCC_{50}) were discarded (81 compound – sequence pairs). **Table 5.1** shows the point mutations present in the 13 mutated sequences as well as the wild type (HXB2 / IIIB reference strain,[27] sequence 1 in the table) and the average pEC_{50} per sequence. A graphic representation of our data set is shown in **Figure 5.1**; a histogram of the pEC_{50} values of all compounds – sequence pairs is available in Supporting **Figure S8**. A sample dataset, the final full model (in the form of a pipeline pilot component), and a protocol to perform PCM are included in the electronic Supporting information.

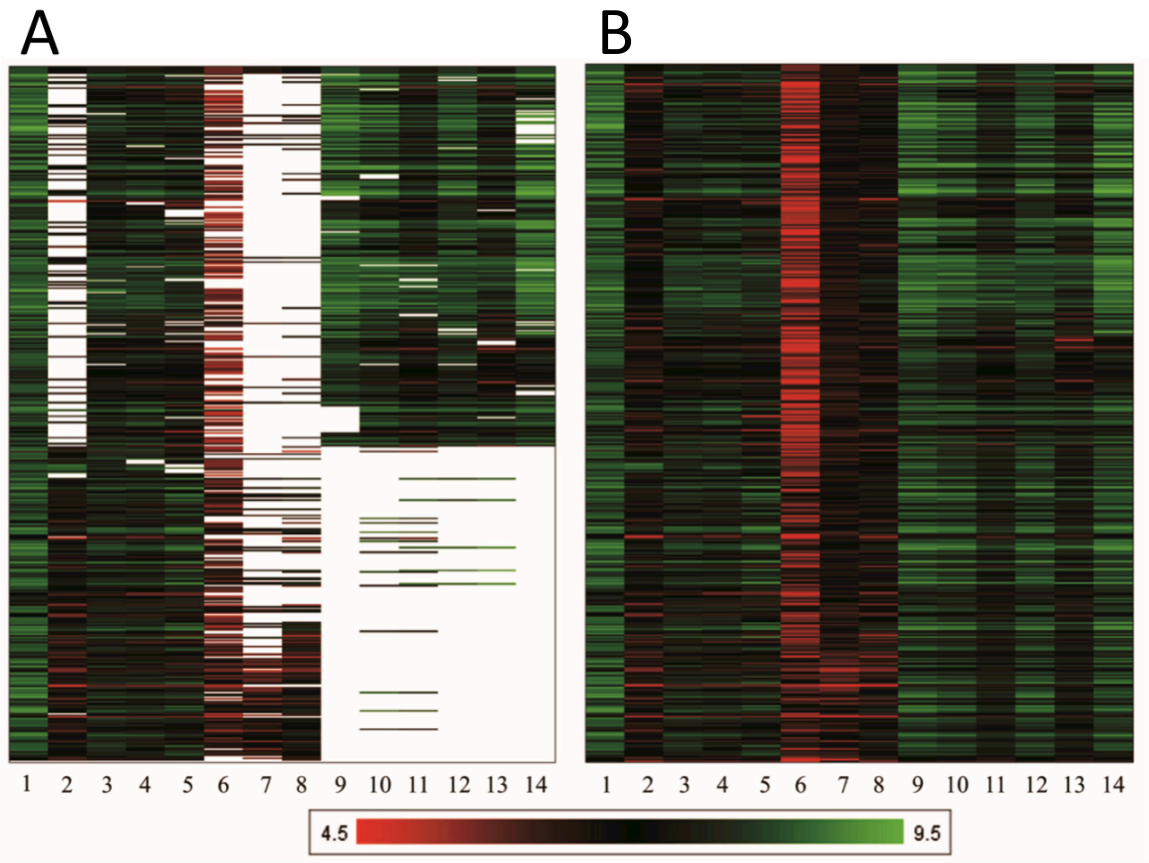


Figure 5.1: Graphical representation of the NNRTI dataset. (A) Our dataset consisted of 451 compounds (in rows) and 14 sequences (in columns) of HIV Reverse Transcriptase with about 60% of the experimental data pairs known. Black indicates a low pEC_{50}, grey indicates a high pEC_{50} and white indicates a missing value. (B) The dataset with the missing pEC_{50} values completed by our model from which 317 experimental validation data points were chosen.

5.3.2 Compound and protein descriptors. All descriptors were calculated in the academic version of PipelinePilot 6.1.5.[28] Ligands were described by Scitegic FCFP_6 circular fingerprints,[29, 30] which have previously been shown to capture a large amount of information with respect to compound bioactivity.[31, 32] FCFP_6 descriptors provide individual substructures and treat these as a feature of a compound. These substructures have a maximal diameter of 6 bonds from a central atom. In the final model we can link these substructures to a change in pEC_{50}.

Sequences were encoded based on the binding site sequence in which each amino acid was represented as a single unique feature. The residues used to define the binding site are shown in **Figure 5.2** (PDB Code 2ZD1, HIV RT bound to Rilpivirine,[33] created with Molsoft ICM version 3.6-h) and **Table 5.1**. The residues used to define the binding site are shown in red and black, where the black residues are the ones that were mutated only in sequence 7.

Table 5.1. Sequence information of the RT sequences in the data set

Sequence / Residue	89	100	101	102	103	106	118	162	169	179	181	188
1	A	L	K	K	K	V	V	S	E	V	Y	Y
2										F	C	
3											C	
4					N						C	
5		I			N							
6		I			N					I	C	
7	S		P	R			I				C	
8			P									
9					N							
10		I										
11								K				L
12			E		N							
13												
14						A			G			

Sequence 1 is wild type RT. The other sequences contained one or multiple point mutations of the binding site residues. The mean pEC_{50} of the compounds tested on the sequence is given as is the standard deviation of the series of compounds tested on the sequence.

The features describing the binding site were obtained by hashing an array of 58 physicochemical properties obtained from the AAindex database;[34] the used indices can be found in Supporting **Table S3**. Finally, both ligand and protein fingerprints were converted to a fixed length array of features which were then used in the modeling. The ligands were converted to a fixed length array of 155 features while the sequences were converted to a fixed length array of 24 features using default Pipeline Pilot settings.

5.3.3 Machine learning. Models were constructed in the academic version of Pipeline Pilot 6.1.5 using the R-statistics package.[28] Support vector machines (SVM) as coded in the e1071 package were used for model creation.[35] Parameters gamma and cost were tuned over an exponential range and epsilon was set at 0.2 as this was the given assay error. It has been shown that setting epsilon to the data error is the optimal value for training.[36] The optimal model was determined using 5-fold cross validation before proceeding to experimental prospective validation of the model. The parameters used for validation were RMSE and R_0^2, although the regular R^2 was also determined it has been shown that R_0^2 provides better reliability.[37, 38] R_0^2, in contrast to R^2, takes into account that the regression line of our model predictions should intersect with the origin. *In silico* validation on trained models was done via learning curves (see Supporting **Figure S9**).

190	203	207	210	211	214	215	219	227	234	245	138 (b)	Mean pEC_{50}	pEC_{50} (sd)	n
G	E	Q	L	R	L	T	K	F	L	V	E	8.3	0.6	451
												6.9	0.7	259
												7.6	0.6	444
												7.5	0.7	443
												7.4	0.8	429
					F						G	6.0	0.6	316
A	V	E	W		F	Y	N			I		6.5	0.6	99
												6.9	0.7	147
												8.3	0.6	222
												7.9	0.7	252
												7.5	0.7	257
												8.0	0.6	242
								C	I			7.4	0.8	244
								L				8.2	0.8	220

n indicates the number of compounds tested on a particular sequence. Residue number 138 is located on the b-chain of RT while the other residues are located on the a-chain.

5.3.4 Prospective Experimental Validation. To assess the prospective capabilities of PCM we used our final model, which was trained on the full data set, to predict the activity of all compounds for which no pEC_{50} value was available. In total 835 data points were subsequently experimentally validated, 317 of these represented novel predictions and 518 were repeat experiments to establish reproducibility of the assay. The 317 novel predictions included 130 compound – sequence pairs that were predicted to differ 2 standard deviations from either compound average (69 compound – sequence pairs), called *compound outliers*, or sequence average (61 compound – sequence pairs), called *sequence outliers*. Therefore we specifically tested the ability of our model to extrapolate outside of the bioactivity space of 'typical' compound – sequence pairs. The techniques used as a benchmark were QSAR, kNN and pEC_{50} scaling. Individual QSAR models were trained for all sequences based on the same FCFP_6 compound descriptors as the PCM model. The kNN models were created based on Jensen *et al*.[39] In total nine kNN models were created. Three models were based on only compound data using 3, 10 or 20 neighbors, three models on only target data using 3, 10 or 14 neighbors and three models based on both compound and target data using 3, 10 or 20 neighbors. In addition we also performed simple interpolation of pEC_{50} values ('scaling'), by assuming that the affinity of a compound on a sequence follows the general trend displayed by a series of other compounds. Thus, if a series of compounds on average has a 0.3 log unit lower pEC_{50} value than on the wild-type sequence, then the new compound is also assigned a 0.3 log unit lower pEC_{50} value.

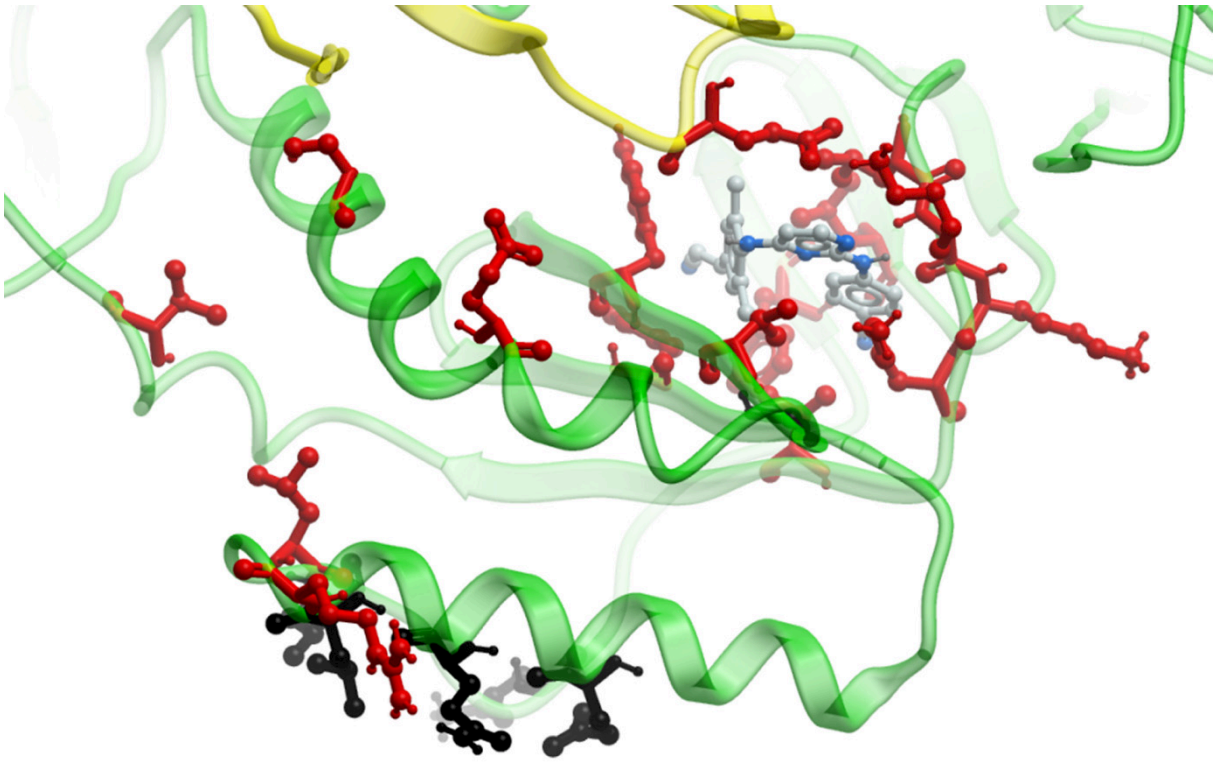


Figure 5.2: The binding site used in our models. PDB X-ray structure 2ZD1 of RT with Rilpivirine, an NNRTI like the analog series we modeled and shown in grey. The residues that were used to identify the binding site (and which were used for activity modeling) are shown in black and red. The black residues are only mutated in sequence 7 (heavy mutant) while the red residues are mutated in more than one sequence. Indicated by a green ribbon is the 'A' chain of HIV RT while the 'B' chain is displayed in yellow.

5.3.5 Antiviral assays. The antiviral activity of different inhibitors was determined in a cell-based HIV-1 replication assay. Here MT4 cells (150,000 cells/ml), stably transformed with an LTR-EGFP reporter gene, were infected with HIV-1 (IIIB, clinical isolates, or site-directed sequence strains; multiplicity of infection MOI = 0.0025) in the presence or absence of different inhibitor concentrations. After three days of incubation, the amount of HIV replication was quantified by measuring the EGFP fluorescence, and expressed as EC_{50} values. The toxicity of inhibitors was determined in parallel on mock-infected MT4 (150,000 cells/ml) cells stably transformed with a CMV-EGFP reporter gene and cultured in the presence or absence of test compound concentrations. After three days of incubation, cell proliferation was quantified by measuring the EGFP fluorescence, and expressed as CC_{50} values (cytotoxicity, 50% inhibitory concentration of cell growth).

5.3.6 Leave-one-sequence-out validation. In order to assess the models' capability to extrapolate between the different sequences we performed 'Leave-one-sequence-out validation' (LOSO). In this approach for each of the 14 sequences all data points of the sequence under consideration are left out of the training set, a model is trained on the remainder and the points left out are used as a test set for that model. Both the RMSE and the R_0^2 values of the validation plot were subsequently determined for each of the 14 sequences in turn.

5.3.7 Model interpretation. To determine the effect of individual residues, for each sequence each residue was mutated back to wild type *in silico*. Subsequently for all compounds the model prediction on the original mutant sequence was compared with the prediction of the model on the *in silico* changed mutant sequence. The difference was interpreted as the change in pEC_{50} induced by that particular residue. From all these changed prediction values the following values were calculated: the average overall, the average per sequence and per individual mutation. This provided the model interpretability. However, changes that led to a 0 value shift in pEC_{50} were removed in the calculation of the average per position since in all cases this was caused by mutation back to wild type of a residue that was already wild type in that particular sequence. In addition the large amount of 0 value shifts lead to a shift of the average towards 0 when it was calculated thereby masking the actual contribution of mutations.

To interpret the influence of compound substructures a slightly different approach was chosen. Here each FCFP_6 feature, corresponding with a particular substructure, was substituted by the feature representing a single Carbon atom ('0'). Since this carbon atom was already present in all compounds, its effect on binding serves as a calibration. Subsequently the full model was used to predict the pEC_{50} of the adapted compound fingerprint lacking a certain substructure on all sequences and compared with the model prediction of the original compound fingerprint on all sequences. From all these changed predictions the average overall and per sequence was calculated for that particular feature, or chemical substructure in this case.

5.3.8 Chemistry. Synthesis of the analog series we used in this work has been described in multiple patents. Compound **1** is listed in patent WO 2007/113256,[40] compound **2** is listed in WO 2008/080965,[41] and compounds **3** and **4** are listed in WO 2008/080964.[42] Compound **6** is listed in patent WO 2008/080965,[41] compounds **5**, **7** and **8** are published in WO2007/113254.[43] In addition the electronic Supporting information to this work contains the structures of 57 of the modeled compounds. For these compounds the full biological activity spectrum on the 14 sequences as we had it available is included along with the pEC_{50} predicted by our model on each of the 14 sequences (Supporting **Table S4** and Supporting archive file).

5.4 Results and Discussion

In the current work PCM was applied to a data set of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), which constitutes one of the major classes of anti HIV drugs on the market. Among these are compounds such as Nevirapine,[44] and Efavirenz,[45] but also novel compounds such as Etravirine,[46] which was approved by the FDA as recently as 2008. Since NNRTIs are allosteric binders they have shown considerably fewer side effects than orthosteric drugs.[47] However they are also known for a quick onset of viral resistance due to the accumulation of resistance associated mutations.[48, 49] Hence, the pharmacological profile of NNRTIs is highly desirable, but their effectiveness is hampered by the onset of resistance. This problem is also encountered in other viral infections like Hepatitis B,[50] Hepatitis C,[51] and Influenza A (H5N1).[52, 53] Resistance can therefore be considered a universal problem when developing a new anti-viral drug. Thus, when new anti-virals are developed it is important that they retain their effectiveness despite the presence of these mutations. To be able to predict the activity of a preclinical drug candidate on an adapted pathogen would be an important contribution to drug discovery. Here we present an application of PCM that can predict drug performance on unknown sequences or adapted pathogens, when staying within model limitations.

5.4.1 Solving the problem of sparse data sets. Our data set contains compounds that inhibit wild type Reverse Transcriptase (RT), but also some that inhibit a number of RT sequences that are highly resistant against inhibition. NNRTIs are allosteric inhibitors of RT, which is illustrated in **Figure 5.2**.[33] Incomplete bioactivity data sets are common in real-world settings, and this study is no exception. The data set is graphically displayed in **Figure 5.1**, with 'blanks' representing unknown data points that we would like to predict using computational methods. A pEC_{50} value was available for approximately 60 % of the data set.

In a preclinical drug discovery setting, the ability to make decisions on a full rather than a sparse matrix increases the likelihood that the best candidate will be selected. This is of vital importance as a SAR table of drug – target interactions does not necessarily show linear relationships, e.g., a substituent in a given compound will not lead to the same increase or decrease in binding on different targets. This is even the case in analog series like our data set.[54, 55] Especially in anti-viral research compounds can display unexpected behavior on the different sequences. It is this behavior we are able to capture and translate into accurate predictions.

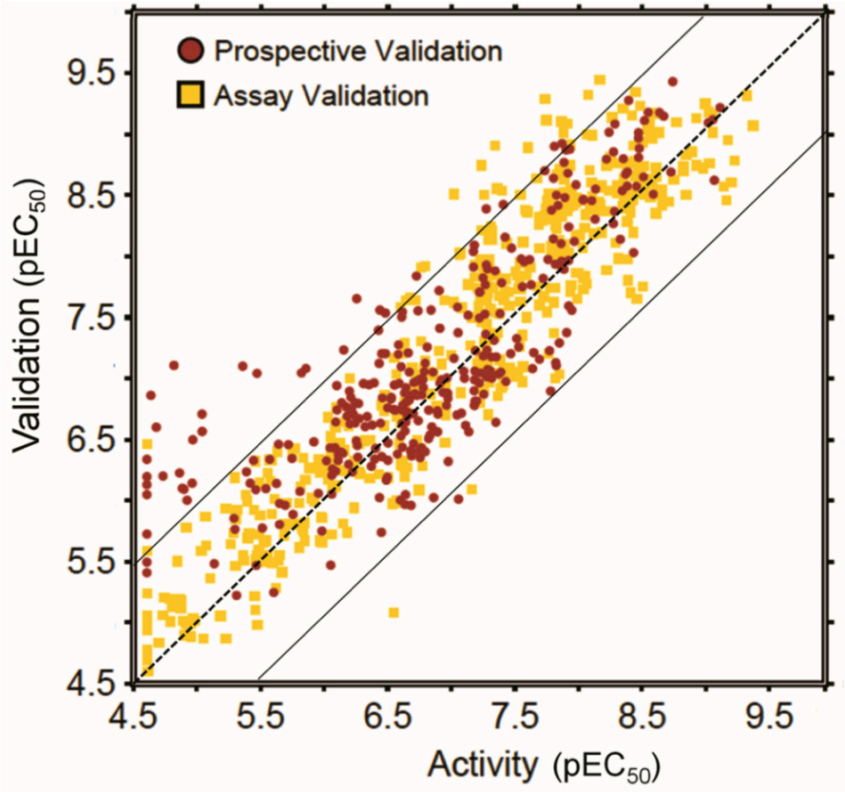


Figure 5.3: Model performance in the prospective experimental validation. Shown are both predictive performance of the model (black dots; R_0^2: 0.69, RMSE: 0.62 log units) and assay reproducibility (grey squares; R_0^2: 0.88, RMSE: 0.50 log units). The continuous lines indicate an error of 1 log unit, while the center dashed line indicates a perfect correlation (see **5.4.2** for further details).

Table 5.2. Performance of different methods in experimental validation

Experiment	Validation	Assay	PCM	pEC_{50} scaling	QSAR	3-NN (both)	3-NN (target)
(Full plot)	R_0^2	0.88	0.69	0.69	0.31	0.38	0.04
(Full plot)	RMSE	0.50	0.62	0.57	0.96	0.90	1.6
(Sequence Outliers)	R_0^2	0.88	0.65	0.52	0.32	0.32	< 0.00
(Sequence Outliers)	RMSE	0.50	0.39	0.52	1.4	0.57	2.3
(Compound Outliers)	R_0^2	0.88	0.56	0.65	0.39	0.30	0.19
(Compound Outliers)	RMSE	0.50	0.65	0.64	0.72	0.87	1.18
(Outliers)	R_0^2	0.88	0.61	0.59	0.36	0.31	< 0.00
(Outliers)	RMSE	0.50	0.52	0.58	1.1	0.72	1.7

PCM and pEC_{50} scaling outperform the other techniques (QSAR and k-Nearest Neighbors) with pEC_{50} scaling having a slight advantage. However, PCM performs better when the selection is narrowed to those compound – sequence pairs that show a pEC_{50} two standard deviations higher or lower than average (shown here as Sequence and Compound Outliers).

5.4.2 Prospective Experimental Model Validation. During model development, learning curves were generated that represent *in silico* validation, in addition we performed Y-scrambling (see Supporting **Figure S9** and **Figure S11**).[56] More importantly, we predicted the pEC_{50} of 317 unknown compound – target pairs. These predictions were subsequently measured experimentally. By predicting first and subsequent experimental validation, we obtain an estimate of model performance when considering novel compound – target pairs. Here we explicitly selected the more 'difficult' compound – target pairs; namely those with predicted pEC_{50} values that are either atypical for the particular compound tested (compound outliers), or for the particular sequence under consideration (sequence outliers). It is trivial to pick the compounds that are always active or always less active – we precisely removed those cases from our prospective testing and focused on the inhibitors that *were predicted to be active against highly resistant sequences, or those which were predicted to be less active against very susceptible sequences.*

3-NN (cmpd)	10-NN (both)	10-NN (target)	10-NN (cmpd)	20-NN (both)	14-NN (target)	20-NN (cmpd)
0.21	0.41	0.21	0.28	0.40	0.21	0.28
1.2	0.90	1.3	1.2	0.90	1.2	1.2
< 0.00	0.32	0.15	0.03	0.32	0.22	0.020
1.68	0.57	1.9	1.7	0.57	1.8	1.7
0.18	0.36	0.49	0.33	0.38	0.54	0.35
1.1	0.86	0.90	0.93	0.86	0.82	0.92
0.08	0.34	0.32	0.18	0.35	0.38	0.19
1.4	0.72	1.4	1.3	0.72	1.3	1.3

Also shown is the average score of each technique on the combined outliers (shown as Outliers). Negative values are denoted as < 0.00.

Figure 5.3 shows the performance of the model, trained on the full data set, in these prospective validation experiments. During cross-validation the full model achieved a Root-Mean-Square Error (RMSE) of 0.38 log units (with a Q^2 of 0.84) while employing 155 ligand features and 23 protein features. When applied to the new *untested* data, the model achieves an RMSE of 0.62 log units and an R_0^2 of 0.69 in the prospective validation. Our model can predict the pEC_{50} of *untested* compound – target pairs with average accuracy of 0.62 log units. This RMSE approaches the reproducibility of the assay which was 0.50 log units (R_0^2 of 0.88).

To benchmark this performance against conventional approaches we applied QSAR modeling, k-Nearest Neighbor (kNN) modeling (using 3, 10 or 20 nearest neighbors and based on compound, target, or compound and target information) and pEC_{50} scaling to the same data set. The results are shown in **Table 5.2**. Here PCM outperforms QSAR and all forms of kNN modeling, while pEC_{50} scaling seems to perform slightly better. More specifically, PCM had an RMSE of 0.62 log units, while kNN showed 0.90 for the best model and scaling performed slightly better with 0.57. The R_0^2 reached 0.69 for PCM, where kNN showed 0.41 and scaling also reached 0.69. Scaling of pEC_{50} values performs second best but this method has two major disadvantages.

When we consider the sequence and compound outliers, PCM outperformed the simple pEC_{50} scaling. It is precisely those data points that points are most interesting in research. The sequence outliers represent the inhibitors that inhibit all present HIV mutants*, candidates to select in lead selection*. The compound outliers can be completely inactive on one particular sequence, *candidates to avoid in lead selection*.

When considering these compounds, the average RMSE was 0.52 for PCM and 0.58 for scaling (R_0^2 was 0.61 for PCM and 0.59 for scaling). A second disadvantage of scaling is that this method cannot be applied to untested viral sequences, whereas PCM can.

Hence, we conclude that the prospective experimental validation confirms the validity of our model, and the applicability of PCM to extrapolate in both ligand (chemical) space and target (biological) space. The other benchmarked techniques, kNN and QSAR, do not accurately capture the compound – target interaction space. The ability to model all these situations with a comparable reliability and the possibility to interpret the model from both a sequence and compound perspective, are the major advantages of PCM over kNN, QSAR and scaling methods.

5.4.3 Neighborhood Behavior in Target Space. In the prospective validation we noticed that a number (15) of data points were predicted inaccurately with an error larger than twice the RMSE, all of which involved sequence 7. Trying to elucidate the reason for this behavior the different sequences were clustered based on similarity (see Supporting **Figure S5** for details), where sequence 7 was found to be most dissimilar to all other sequences, containing a rather large number of 13 point mutations (for sequence information see **Table 5.1**). In addition, it is the sequence with the smallest number of tested compounds in the training set, thus diminishing the number of compound – target pairings that were used in the model directly (without extrapolation from neighboring sequences), thereby rationalizing the large number of false predictions on this sequence. In order to systematically investigate the dependence of the prediction error on the sequence similarity we plotted the prediction error as a function of the average Tanimoto similarity of the individual sequences to the rest of the training set (shown in **Figure 5.4**). A correlation between model error and average sequence similarity is observed. The prediction error increases when the average similarity decreases between the sequence, for which the compound activity is predicted, and the rest of the training set sequences.

This observation is an extension to the 'Molecular Similarity Principle' which states that similar compounds have similar properties, and in this work we are able to show that this paradigm also holds in biological space *where similar sequences show similar ligand binding abilities*; a concept we are now able to quantify numerically.This extends recent work on 'applicability domains' and 'activity cliffs' by also taking the biological target side into account.[54, 55, 57, 58] An improved distance measure is an ensemble of both the distance between the compounds (Supporting **Figure S10**), as has been previously shown,[16] and the distance on the *target side*. To our knowledge this neighborhood behavior in target space has not been previously shown but is a natural extension of the *chemical similarity principle* to this new technique.

The dependence of model accuracy on the average distance provides a useful tool to set the model applicability domain. As this distance is a property dependent on the training set on one hand and the unknown compound – target pair on the other hand it can be measured before any model prediction is made. Furthermore, neighborhood behavior can determine beforehand if the model is capable of predicting pEC_{50} changes on a previously untested genotype. This approach ensures that only model predictions with certain accuracy are used and that those that do not meet this accuracy are disregarded. For our current model this accuracy is defined as an error in pEC_{50} prediction (**Figure 5.4**) but this can also be an error in pKi value or any other value the model is trained to predict.

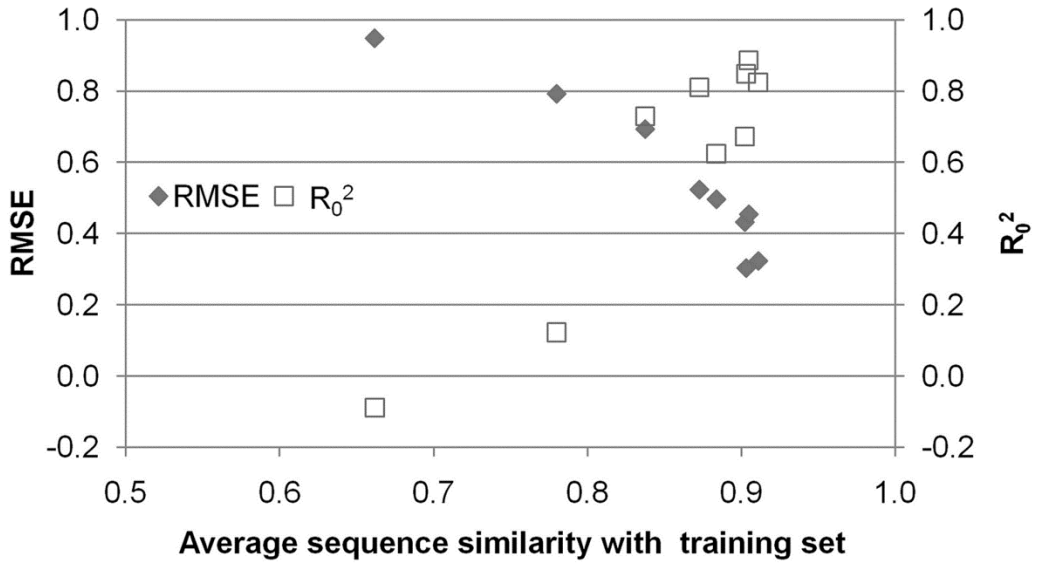


Figure 5.4: Extension of the applicability domain to target space. Prediction error (measured as R_0^2 and RMSE in log units) versus the average similarity of the sequence to the rest of the training set, extending the 'Molecular Similarity Principle' to biological space which is of crucial relevance in PCM. It can be seen that a higher similarity to the training set leads to more accurate predictions. Still, predictions on the most dissimilar sequence have an average error of less than 1 log unit.

5.4.4 How to anticipate bioactivity for novel protein targets? As we have shown that our model displays neighborhood behavior in target space, here we explore the possibility to use such a model in extrapolation. Conventionally a fraction of the full data set is left out from the training set when testing bioactivity models. The ability of the model to make predictions for the previously untested ('novel') compounds is taken as a predictor of model performance. In our case, we not only extrapolated in *compound space*, but also in *sequence space*. Hence, in order to confirm the ability of the model to extrapolate the activity of compounds to related sequences we performed a 'Leave-one-sequence-out' experiment (LOSO). Here, it emulates the prediction of inhibitor activity for a virus with a *not previously* encountered RT sequence, based solely on bioactivity measurements against other sequences in the data set. While applied to enzyme mutants here, the concept is generally applicable and the authors are currently investigating its performance on other target families such as GPCRs.

Figure 5.5 shows the R_0^2 and RMSE of the LOSO validation experiment (see Supporting **Figures S1** and **S2** for additional information including error depending on similarity to the training set in sequence space). We observe that for 13 of the 14 sequences an RMSE of less than one log unit was obtained, and five of the 14 sequences even yielded an RMSE of less than 0.5 log units, which is the order of magnitude of assay reproducibility. Most interestingly, the model was able to use the information contained in all the different mutant sequences and predict the affinity of the compounds on the wild type sequence (Sequence 1). This means that this LOSO-1 model was able to deconvolute the individual contributions of the different mutants. Furthermore other LOSO models were also capable of predicting activity on unknown sequences 2-5 and 9-14. All these sequences contain very different mutations (**Table 5.1**). Finally, the LOSO-7 model was able to predict the pEC_{50} of the compounds on the "heavy" mutant (sequence 7), which contains a total of 13 point mutations. These findings underline the ability of PCM to extrapolate in target space, which supports the application of the technique to predict compound activity on different mutants.Our results show that the model is indeed able to predict the pEC_{50} values of a known compound on an unknown sequence. However it should be noted that PCM performs lesser in 2 of the 14 mutants considered. The first of the two exceptions is sequence 6, which can be seen as a singleton since one of the mutations it carries (E138G) is only present in this particular sequence. Therefore, bioactivity prediction based on other sequences that do not carry this mutation is not straightforward. It was already known from the full model interpretation (see **5.4.6**) that the impact of this mutant was underestimated. The LOSO-6 model correctly predicted that all compounds have a lower activity on this sequence (giving rise to a small RMSE value); however, the ranking among compounds is not very accurate (explaining the low R_0^2).

The second sequence where our model underperforms is sequence 8. This case is different from the previous one: while the model is able to correctly rank the compounds, leading to an acceptable R_0^2 of about 0.6, it consistently overpredicts the activity of compounds, leading to a high RMSE. In sequence 8 at position 101 a positively charged lysine residue is replaced with a proline, which likely induces conformational changes to the backbone of the protein. K101P is present as a single mutation in sequence 8, as well as in combination with a total of 12 other mutations in sequence 7 (which is then the only sequence that contains K101P in the training set when sequence 8 is left out for LOSO validation).

Hence, the model likely underestimates the impact of the K101P mutation due to the large number of other mutations present in sequence 7; it is not able to deconvolute the impact of every single mutation properly. In this case, while the model would overestimate drug activity on this particular target, in a computer-aided compound selection setting the correct candidate would still be identified due to the accurate ranking of the compounds by the model.

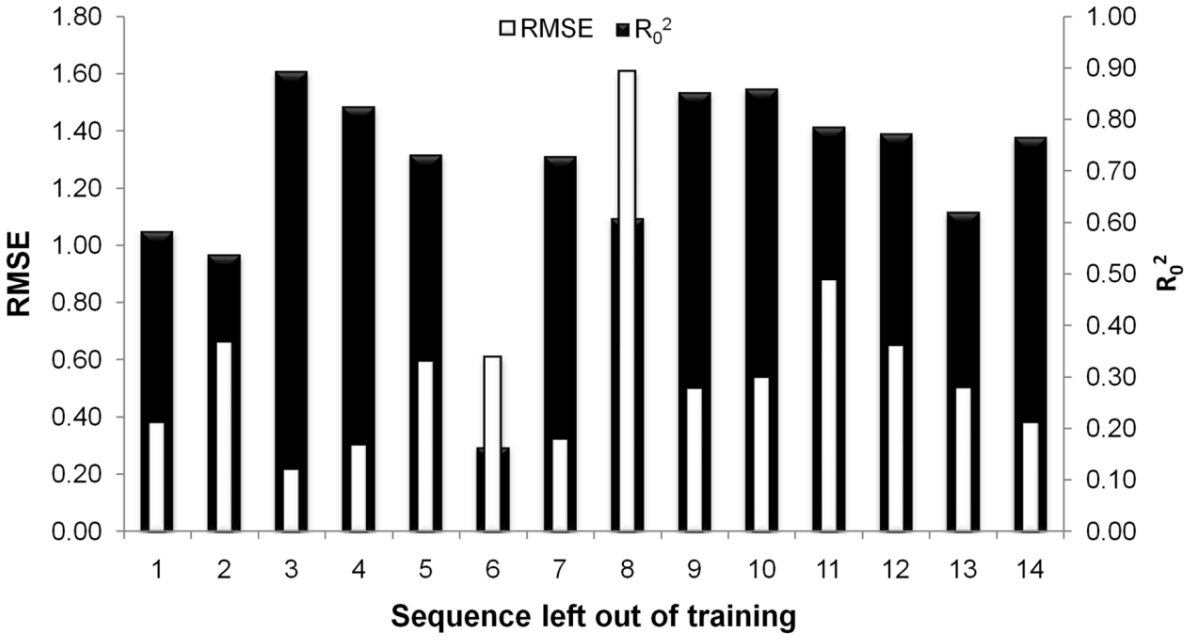


Figure 5.5: Performance of PCM in leave-one-sequence-out experiments. Performance was measured by R_0^2 and the RMSE in log units. The number below the bar corresponds to the sequence left out of the training set.

5.4.5 Model performance in relation to chemical structure. To get an idea of the ability of the LOSO models' ability to predict individual compounds we have shown representative examples of compounds either predicted accurately or inaccurately in **Figure 5.6**. The activities of compounds **1** and **2** on the different sequences were correctly predicted by our leave-one-sequence-out models, while compounds **3** and **4** were predicted inaccurate (see Supporting **Figures S3** and **S4** for the predictions and experimental activities of the compounds discussed here). Please note that we made predictions challenging as we used LOSO models to predict the activity of compounds on the sequences left out of the training of these exact LOSO models. It is therefore a realistic emulation of a preclinical PCM application. From the data we conclude that inaccurately predicted compounds have a large functional group in the 4 position of the pyrimidine ring. Apparently the LOSO models are unable to capture this information correctly. When we further analyzed the individual predictions on the different mutants (shown in Supporting **Figure S4**), we noticed that the compounds are predicted accurately (error < 0.5 log units) on the majority of the sequences. The underperformance for compound **3** is caused by overprediction on sequence 8 (2.7 log units, carrying K101P) and the underperformance of compound **4** by underprediction on sequence 2 (carrying V179F).

To explain this behavior, we need to consider the ligand binding mode. A shared binding mode of all compounds is very likely since i), they all share a common chemical scaffold and ii), NNRTIs are known to have a highly homologous binding mode.[59] Hence we can correlate these mispredictions with the protein structure. The substitution position on the pyrimidine ring in the compounds corresponds to the location of residues L100, K101 and V179 in crystal structure 2ZD1. It is known that these residues mutate easily and sequences carrying point mutations in this location are present in our training set.[60] When studying the crystal structure, it can be seen that the side chains of these residues are likely to hamper the binding of compounds with a large functional group in position 4.

Hence, we propose the presence of a large functional group on position 4 on the pyrimidine ring to be a predictor of insufficient model performance in combination with mutations on positions 101 and 179 in the protein. While compounds *with* a large substituent on the 4 position are accurately predicted on targets *not carrying* mutations on positions 101 and 179, they tend to be predicted inaccurately when the targets *are mutated* in these positions. However, if no large substituent is present on the 4 position, compounds *are* predicted accurately on targets carrying mutations on positions 101 and 179. Using this knowledge we can define an applicability domain when applying this particular model, but this finding should also be taken as a warning when using any PCM model to predict the affinity of known compounds on unknown sequences.

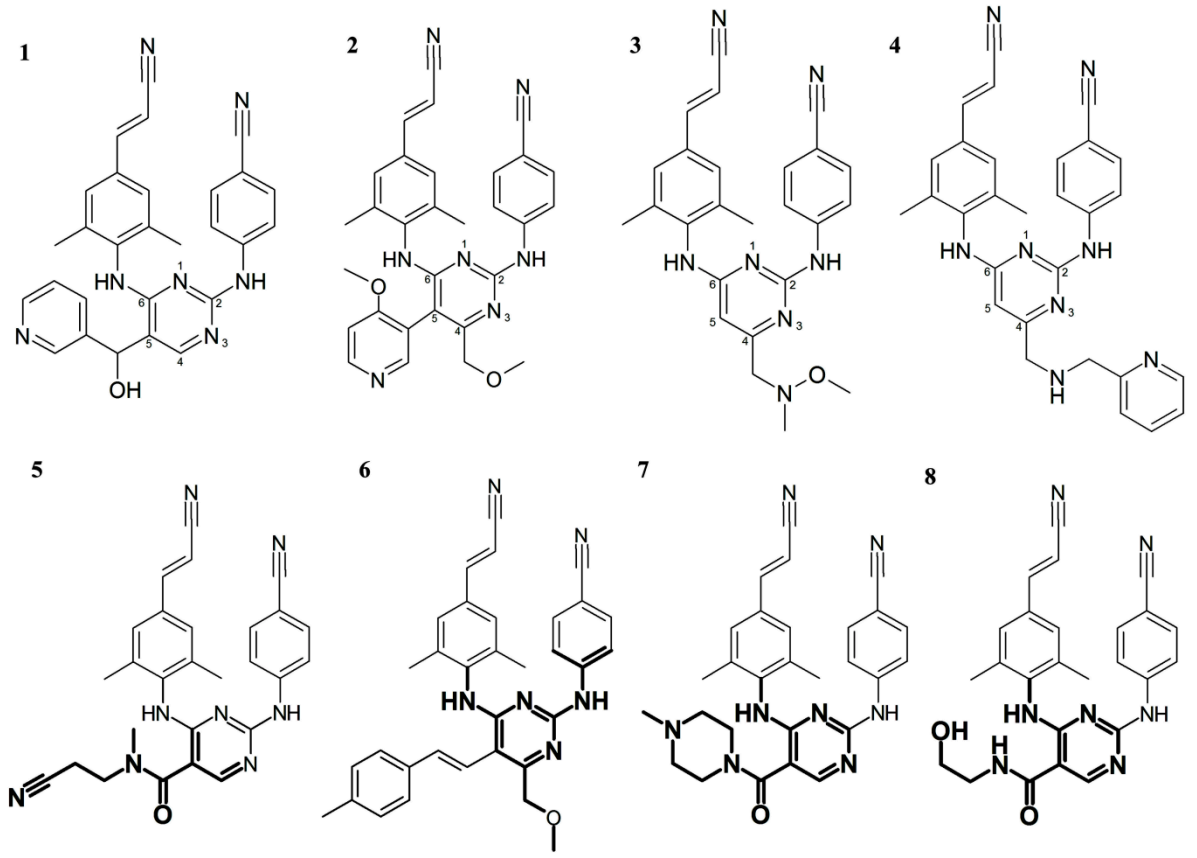


Figure 5.6: Example structures that where included in the model. A selection of both compounds containing accurately (**1**,**2**) and inaccurately modeled chemistry (**3**,**4**). Also shown are compounds containing a substructure positively correlated with pEC_{50} (**5**,**6**) and compounds containing a substructure negatively correlated with pEC_{50} (**7**,**8**). In the upper part, shown are sample compounds that were accurately predicted using the LOSO models, a low RMSE (**1**, RMSE was 0.22 log units, R_0^2 was 0.96) and a high R_0^2 (**2**, R_0^2 was 0.89 , RMSE was 0.26). Secondly, sample compounds that were predicted inaccurately using the LOSO models, a high RMSE (**3**, RMSE was 1.12 log units, R_0^2 was -0.10) and a low R_0^2 (**4**, R_0^2 was -3.66, RMSE was 0.61). The lower part shows the 17th best substructure (**5**) and the 30th best substructure (**6**). Conversely the 3rd worst substructure (**7**) and the 4th worst substructure (**8**) are depicted.

5.4.6 Model based interpretation of mutations. After completion of the full validation, the final full model was subsequently interpreted to explain the differences in activity of the compounds on the individual mutants. This is the model we used to perform the experimental validation and not one of the LOSO models. Firstly we focused on the sequence side of the model (**Table 5.1** and for a multiple sequence alignment see Supporting **Figure S6**). By correlating model predictions of all compounds on each mutant sequence, where all amino acids were replaced in turn by their wild type counterpart on that particular position, an overview of the variation present at all individual residues was created (**Figure 5.7**). The full model explains the lowered activity (pEC_{50}) on mutants mainly by mutations at residues 100, 101, 103, 162, 179, 181, 188, 227, 234 and slightly by residue 138 on the 'b' chain. Interestingly the model interprets mutations at positions 89, 102, 118, 190, 203, 207, 210, 215, 219, 245 to actually slightly increase compound activity. Residues 106, 169 and 214 have little influence in this model and residue 211 does not seem to contribute at all (Supporting **Figure S7** shows the variation in pEC_{50} caused by individual mutations).

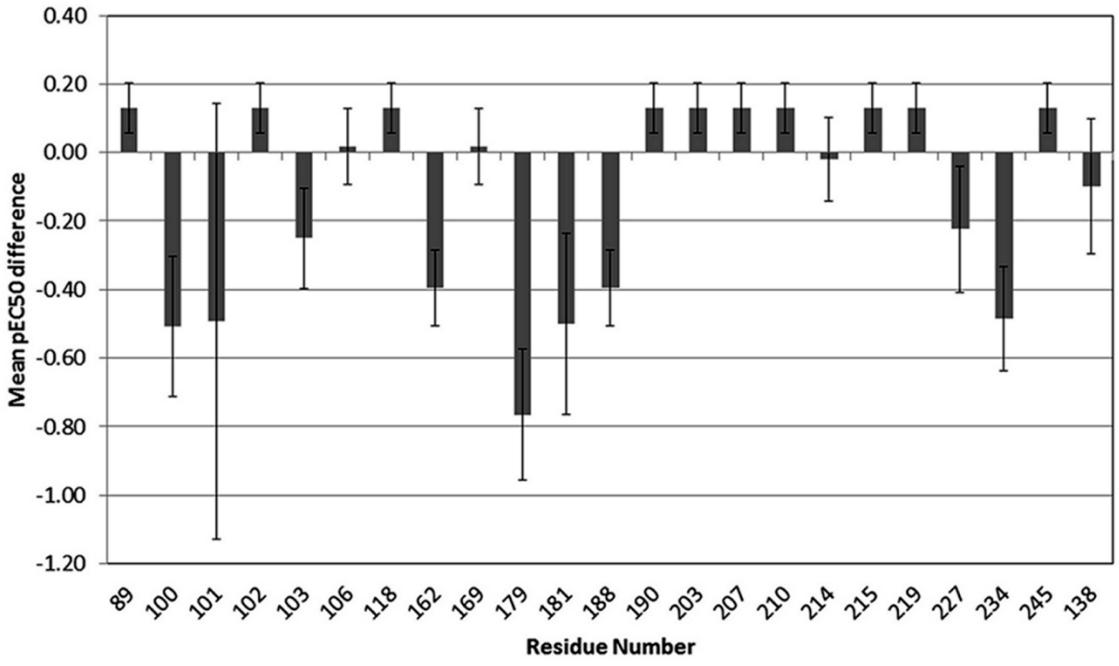


Figure 5.7: Overview of the contribution of mutations present at all individual residue positions. The full model explains the lowered activity (pEC_{50} in log units) on mutants mainly by mutations at residues 100, 101, 103, 162, 179, 181, 188, 227, 234 and residue 138 on the 'b' chain.

The main contributor to lowered affinity appears to be residue 179 rather than residue 181, while the latter is known to lead to NNRTI cross resistance.[60] Supporting this finding, it is known that clinically used Etravirine, a compound similar to but not included in our data set, is also sensitive to mutations at position 179 (V179D, V179F, and V179T).[60] Secondly, the influence of mutations at position 101 differs widely indicated by the high standard deviation. This is caused mainly by the K101P mutant, which causes a large decrease in pEC_{50} for some compounds and very little for others, depending on the chemistry (see **5.4.7**). The K101E mutant overall has little influence. This interpretation is in line with the results from the LOSO experiments mentioned above where both residues 101 and 179 were identified as having a high impact on model reliability.

5.4.7 Model based interpretation of ligand substructures. Similar to the mutation interpretation, the full model was interpreted to elucidate the average contribution of individual compound substructures to changes in activity (**Figure 5.8**). Here the substructure that was being investigated was replaced by a single carbon atom in all compounds and the subsequent model predictions were compared to the model predictions for the original compounds. **Figure 5.8** shows the contribution of the substructures after a selection was made to only use substructures that occur in more than one compound in order to lower a bias of a continuously active compound (resulting in 1068 of 2546 substructures). In the figure some examples of substructures that improve activity and substructures that decrease activity are shown. For a full table with the top 15 best and top 15 worst substructures, please see the Supporting **Table S1** and Supporting **Table S2**. **Figure 5.6** (compounds **5,6**) shows two examples of compounds that contain a substructure, which has been highlighted, that the model predicts to lead to a good activity and two examples of compounds (**7,8**) that contain a substructure that the model predicts to have a negative effect on activity. The effect is expressed as an average increase or decrease of all compounds containing that substructure and their activities on all sequences. Compound **5** contains the 17^{th} best substructure, leading to an average increase in pEC_{50} of 0.14, and compound **6** contains the 30^{th} best substructure, leading to an average increase in pEC_{50} of 0.12. Therefore these substructures constitute chemistry that is optimally contained in the compound. Compound **7** contains the 3^{rd} worst substructure, leading to an average decrease in pEC_{50} of 0.23, and compound **8** contains the 4^{th} worst substructure, leading to average decrease in pEC_{50} of 0.19. These two substructures constitute chemistry that is rather avoided in possible drug candidates.

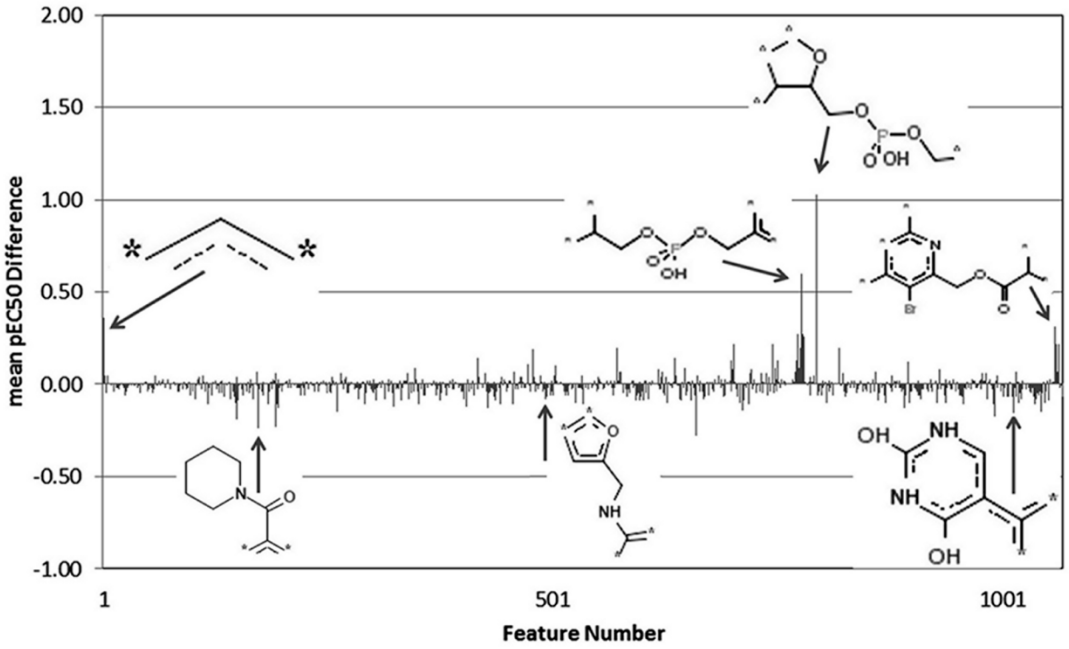


Figure 5.8: Overview of the contribution of the different chemical substructures. Substructures occurring only in a single compound have been removed and the remaining substructures have been numbered sequentially. Several substructures have been visualized and linked to their position on the overview.

5.4.8 Application of PCM in preclinical drug research. We have shown that PCM can be applied in a preclinical setting to predict the resistance profile of compounds. In addition we can interpret our final full model and identify favorable and unfavorable substructures, providing insights that can be used in compound design. For this data set we can conclude that the mutations in the protein sequences have a larger impact on pEC_{50} values than the compound substructures. However, we have also shown that the substructures still possess a significant influence, as PCM was the only technique that was able to combine target and substructures information.

Combining all our results from prospective experimental validation, the LOSO and the model interpretation, we feel confident that our model can be used to estimate the activity of previously untested compound – sequence pairs, with the main limitation (which can be quantified) being the similarity of the target protein. This opens the door for models that are able to predict the changes in activity of different compounds on clinical isolates obtained from patients.[13] PCM can thereby serve as a modeling tool to predict the activity for untested compound – isolate pairs before any assay measurement is performed, providing a quick guidance to medicinal chemists in the development of drugs based on their expected resistance profile.

An example of this application is given in **Table 5.3** and **Table 5.4**, listing which inhibitors show best activity against a particular sequence and for which inhibitors resistance would be expected. (Supporting **Table S4** lists a total of 57 compounds and their activity on the 14 sequences) Likewise, our method also aids in the development of drugs that have a broad inhibition profile. Noteworthy is that eight of the 28 specific compound - sequence predictions are untested compound – sequence pairs, which underlines the value of PCM to extrapolate to untested drug – sequence pairings, a feature not possible to achieve in conventional (single-target based) bioactivity modeling.

Table 5.3. Best performing compounds (per sequence and overall)

Sequence	Compound with highest pEC_{50}	Activity (pEC_{50})	Full Model (pEC_{50})	Difference (Activity and Model)
All	326	8.39 (± 0.61)	8.53 (± 0.73)	0.14
1	365	9.16	9.55	0.39
2	221	8.19	8.38	0.19
3	79	8.71	8.81	0.10
4	321	8.83	8.79	0.04
5	321	9.12	8.73	0.39
6	221	8.01	7.93	0.08
7	364	untested	7.50	n/a
8	221	untested	8.42	n/a
9	365	untested	9.43	n/a
10	326	untested	9.23	n/a
11	151	9.05	8.86	0.19
12	321	untested	9.29	n/a
13	100	9.06	8.87	0.19
14	79	9.51	9.62	0.11
			Average	0.18

Overview of the compounds with the highest pEC_{50} as obtained from the model. Shown are the pEC_{50} values differentiated over all sequences (all) or per sequence. Also shown is the standard deviation of the distribution over all sequences used to calculate this mean value. It should be noted that compound **326** was not tested on sequences 9, 10 and 14, illustrating the importance of extrapolating in bioactivity space.

Table 5.4. Worst performing compounds (per sequence and overall)

Sequence	Compound with Lowest pEC_{50}	Activity (pEC_{50})	Full Model (pEC_{50})	Difference (Activity and Model)
All	109	5.85 (±0.54)	5.82 (±0.66)	0.03
1	248	6.09	6.01	0.08
2	109	untested	4.87	n/a
3	422	untested	5.78	n/a
4	84	5.84	5.67	0.17
5	84	5.65	5.54	0.11
6	109	4.60	4.06	0.54
7	439	5.01	5.20	0.19
8	84	4.74	5.20	0.46
9	248	untested	5.96	n/a
10	181	5.82	6.01	0.19
11	181	5.42	5.61	0.19
12	109	5.90	6.09	0.19
13	181	5.11	5.29	0.18
14	181	5.62	5.81	0.19
			Average	0.21

Overview of the compounds with the lowest pEC_{50} as obtained from the model. Shown are the pEC_{50} values differentiated over all sequences (all) or per sequence. Also shown is the standard deviation of the distribution over all sequences used to calculate this mean value. It should be noted that compound **109** was not tested on sequences 2, 7 and 8, illustrating the importance of extrapolating in bioactivity space.

5.5 Conclusion

In this work we have shown how to incorporate personalized data (specific viral mutants) as a tool to select optimal candidates in drug development by using proteochemometric modeling in combination with a large scale experimental validation of inhibitors of HIV Reverse Transcriptase. While applied here to NNRTIs, PCM is a universally applicable method since all it requires is the sequence of a target of interest and structures of ligands. These two prerequisites are something that is available in any preclinical drug research project. We employed a prospective validation of 317 new experimental data points and a new type of 'leave one sequence out' validation (representing the case of a previously untested virus genotype). We were able to predict which compounds are best for a particular HIV RT sequence (with high accuracy in 12 out of the 14 sequences in the data set). We established that distance in biological (target) space is tightly correlated with prediction performance, enabling us to judge where the model likely succeeds, and where it may fail. Hence, in this work, we present a real-world scenario of HIV drug development, make practical steps towards drug design tailored towards specific patients, and we aim to extend this to other target families in the future.

5.6 Acknowledgements

G. v. W. would like to thank Tibotec BVBA for generously providing the data set.

5.7 Supporting Information

Additional tables (Supporting **Table S1** – **Table S4**), figures (**Figure S1** – **Figure S11**), the final model and 57 of the here modeled compounds with biological activity are available online. A protocol to be run in pipeline pilot to apply this model and perform PCM is also available online. These materials are available online at www.gjpvanwesten.nl.

5.8 References

1. J.C. Venter, M.D. Adams, et al.; *The Sequence of the Human Genome.* Science; *2001*. **291** (5507): 1304-1351.
2. K.A. Frazer, S.S. Murray, et al.; *Human genetic variation and its contribution to complex traits.* Nat. Rev. Genet.; *2009*. **10** (4): 241-251.
3. A.L. Hopkins and C.R. Groom; *Opinion: The druggable genome.* Nat. Rev. Drug Discovery; *2002*. **1** (9): 727.
4. A.P. Russ and S. Lampel; *The druggable genome: an update.* Drug Discov. Today; *2005*. **10** (23-24): 1607-1610.
5. K. Hambly, J. Danzer, et al.; *Interrogating the druggable genome with structural informatics.* Mol. Diversity; *2006*. **10** (3): 273-281.
6. J. Woodcock; *The Prospects for "Personalized Medicine" in Drug Development and Drug Therapy.* Clin. Pharmacol. Ther.; *2007*. **81** (2): 164-169.
7. L. Mancinelli; *Pharmacogenomics: the promise of personalized medicine.* The AAPS Journal; *2002*. **2** (1): 29.
8. C. Wang, Y. Mitsuya, et al.; *Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance.* Genome Res.; *2007*. **17** (8): 1195-1201.
9. E.R. Mardis; *The impact of next-generation sequencing technology on genetics.* Trends Genet.; *2008*. **24** (3): 133-141.
10. K. Hertogs, M.-P. de Bethune, et al.; *A Rapid Method for Simultaneous Detection of Phenotypic Resistance to Inhibitors of Protease and Reverse Transcriptase in Recombinant Human Immunodeficiency Virus Type 1 Isolates from Patients Treated with Antiretroviral Drugs.* Antimicrob. Agents Chemother.; *1998*. **42** (2): 269-276.

11. N. Beerenwinkel, B. Schmidt, et al.; *Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype.* Proc. Natl. Acad. Sci. U. S. A.; *2002*. **99** (12): 8271-8276.
12. N. Beerenwinkel, T. Lengauer, et al.; *Methods for optimizing antiviral combination therapies.* Bioinformatics; *2003*. **19** (suppl 1): i16-i25.
13. H. Vermeiren, E. Van Craenenbroeck, et al.; *Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling.* J. Virol. Methods; *2007*. **145** (1): 47-55.
14. A. Altmann, T. Sing, et al.; *Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from HIV genotype.* Antiviral Therapy; *2009*. **14**: 273.
15. A. Altmann, M. Däumer, et al.; *Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database.* J. Infect. Dis.; *2009*. **199** (7): 999-1006.
16. M.A. Johnson and G.M. Maggiora; *Concepts and Applications of Molecular Similarity; 1990*; New York: John Wiley & Sons.
17. D.E. Patterson, R.D. Cramer, et al.; *Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors.* J. Med. Chem.; *1996*. **39** (16): 3049-3059.
18. A. Bender and R.C. Glen; *Molecular similarity: a key technique in molecular informatics.* Org. Biomol. Chem.; *2004*. **2**: 3204-3218.
19. E. Jacoby; *Chemogenomics : knowledge-based approaches to drug discovery; 2006*: Imperial College Press, London.
20. S. Garland and D. Gloriam; *Methods for the Successful Application of Chemogenomics to GPCR Drug Design.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1870-2009.
21. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
22. M. Lapins and J.E.S. Wikberg; *Proteochemometric Modeling of Drug Resistance over the Mutational Space for Multiple HIV Protease Variants and Multiple Protease Inhibitors.* J. Chem. Inf. Model.; *2009*. **49** (5): 1202-1210.
23. J. Meslamani and D. Rognan; *Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel.* J. Chem. Inf. Model.; *2011*. **51** (7): 1593–1603.

24. G.J.P. Van Westen, J.K. Wegner, et al.; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.

25. R. Morphy and Z. Rankovic; *Designed Multiple Ligands. An Emerging Drug Discovery Paradigm.* J. Med. Chem.; *2005*. **48** (21): 6523-6543.

26. R. Morphy and Z. Rankovic; *Fragments, network biology and designing multiple ligands.* Drug Discov. Today; *2007*. **12** (3-4): 156-160.

27. B.T. Korber, B.T. Foley, et al. *Numbering Positions in HIV Relative to HXB2CG*. *1998*.

28. Accelrys Software Inc *Pipeline Pilot Student Edition* Scitegic Version 6.1.5

29. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.

30. R.C. Glen, A. Bender, et al.; *Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME.* IDrugs; *2006*. **9** (3): 199 - 204.

31. A. Bender, H.Y. Mussa, et al.; *Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance.* J. Chem. Inf. Comput. Sci.; *2004*. **44** (5): 1708-1718.

32. A. Bender, J.L. Jenkins, et al.; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.

33. K. Das, J.D. Bauman, et al.; *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: Strategic flexibility explains potency against resistance mutations.* Proc. Natl. Acad. Sci. U. S. A.; *2008*. **105** (5): 1466-1471.

34. S. Kawashima, H. Ogata, and M. Kanehisa; *AAindex: Amino Acid Index Database.* Nucleic Acids Res.; *1999*. **27** (1): 368-369.

35. E. Dimitriadou, K. Hornik, et al. *Misc Functions of the Department of Statistics (e1071)* TU Wien 2006 1.5-15

36. V. Vapnik; *The Nature of Statistical Learning; 1995*; New York: Springer.

37. A. Tropsha, P. Gramatica, and Vijay K. Gombar; *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models.* QSAR Comb. Sci.; *2003*. **22** (1): 69-77.

38. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

39. B.F. Jensen, C. Vind, et al.; *In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-Nearest Neighbor and Extended Connectivity Fingerprints,*

Including Structural Fragment Analysis of Inhibitors versus Noninhibitors. J. Med. Chem.; *2007*. **50** (3): 501-511.

40. J.E.G. Guillemont , C.I. Mordant, and B.A. Schmitt. *HIV Inhibiting 5-(Hydroxymethylene and Aminomethylene) Substituted Pyrimidines* W.I.P. Organization *2007* WO/2007/113256
41. J.E.G. Guillemont , C.I. Mordant, and B.A. Schmitt. *HIV Inhibiting 5,6-Substituted Pyrimidines* W.I.P. Organization *2008* WO/2008/080965
42. J.E.G. Guillemont and C.I. Mordant. *HIV Inhibiting 6-Substitited Pyrimidines* W.I.P. Organization *2008* WO/2008/080964
43. J.E.G. Guillemont , M. Paugam, and B. Delest, François, Marie. *HIV Inhibiting 5-Amido Substituted Pyrimidines* W.I.P. Organization *2007* WO/2007/113254
44. V.J. Merluzzi, K.D. Hargrave, et al.; *Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor.* Science; *1990*. **250** (4986): 1411-1413.
45. S. Young, S. Britcher, et al.; *L-743, 726 (DMP-266): a novel, highly potent nonnucleoside inhibitor of the human immunodeficiency virus type 1 reverse transcriptase.* Antimicrob. Agents Chemother.; *1995*. **39** (12): 2602-2605.
46. K. Andries, H. Azijn, et al.; *TMC125, a Novel Next-Generation Nonnucleoside Reverse Transcriptase Inhibitor Active against Nonnucleoside Reverse Transcriptase Inhibitor-Resistant Human Immunodeficiency Virus Type 1.* Antimicrob. Agents Chemother.; *2004*. **48** (12): 4680-4686.
47. J.C. Wu, T.C. Warren, et al.; *A novel, dipyridodiazepinone inhibitor of HIV-1 reverse transcriptase acts through a nonsubstrate binding site.* Biochemistry; *1991*. **30** (8): 2022-2026.
48. D.D. Richman; *Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy.* J. Virol.; *1994*. **68** (3): 1660.
49. D.V. Havlir, S. Eastman, et al.; *Nevirapine-resistant human immunodeficiency virus: kinetics of replication and estimated prevalence in untreated patients.* J. Virol.; *1996*. **70** (11): 7894-7899.
50. S.N.S. Ahmed, D. Tavan, et al.; *Early detection of viral resistance by determination of hepatitis B virus polymerase mutations in patients treated by lamivudine for chronic hepatitis B.* Hepatology; *2000*. **32** (5): 1078-1088.
51. J.-M. Pawlotsky, G. Germanidis, et al.; *Interferon Resistance of Hepatitis C Virus Genotype 1b: Relationship to Nonstructural 5A Gene Quasispecies Mutations.* J. Virol.; *1998*. **72** (4): 2795-2805.

52. A. Moscona; *Global Transmission of Oseltamivir-Resistant Influenza.* N. Engl. J. Med.; *2009*. **360** (10): 953-956.

53. M.D. de Jong, T.T. Thanh, et al.; *Oseltamivir Resistance during Treatment of Influenza A (H5N1) Infection.* N. Engl. J. Med.; *2005*. **353** (25): 2667-2672.

54. R. Guha and J.H. VanDrie; *Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs.* J. Chem. Inf. Model.; *2008*. **48** (3): 646-658.

55. M.T. Sisay, L. Peltason, and J.r. Bajorath; *Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure-Activity Relationships in Analog Series.* J. Chem. Inf. Model.; *2009*. **49** (10): 2179-2189.

56. L. Eriksson, J. Jaworska, et al.; *Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs.* Environ. Health Perspect.; *2003*. **111** (10): 1361-1375.

57. H. Dragos, M. Gilles, and V. Alexandre; *Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models.* J. Chem. Inf. Model.; *2009*. **49** (7): 1762-1776.

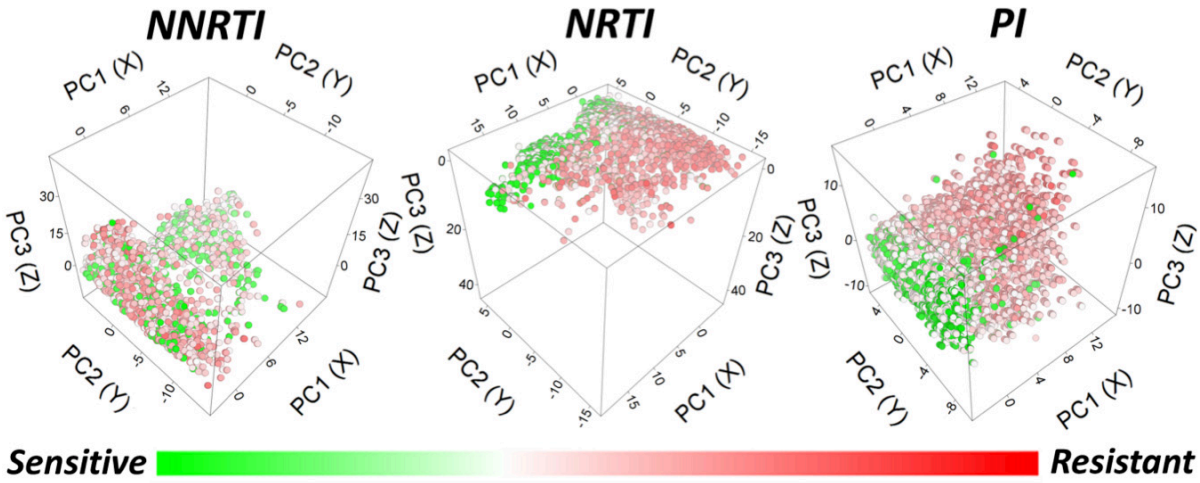
58. J.L. Medina-Franco, K. Martinez-Mayorga, et al.; *Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs.* J. Chem. Inf. Model.; *2009*. **49** (2): 477-491.

59. J. Ren, R. Esnouf, et al.; *High resolution structures of HIV-1 RT from four RT-inhibitor complexes.* Nat. Struct. Mol. Biol.; *1995*. **2** (4): 293-302.

60. V.A. Johnson, F. Brun-Vézinet, et al.; *Update of the drug resistance mutations in HIV-1: December 2010.* Topics in HIV Medicine; *2010*. **18** (5): 156-163.

Chapter 6

Personalized HIV Treatment Regimen Prediction Employing Proteochemometric Models Generated From Antivirogram Data



G.J.P. Van Westen, A. Hendriks, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender;

(Manuscript submitted)

Contents

6.1 Abstract

Infection with HIV cannot currently be cured; however it can be controlled by combination treatment with multiple anti-retroviral drugs. Given different viral genotypes for virtually each individual patient, the question now arises which drug combination to use to achieve effective treatment. With the availability of viral genotypic data and clinical phenotypic data, it has become possible to create computational models able to predict an optimal treatment regimen for an individual patient. Current models are based only on sequence data derived from viral genotyping; chemical similarity of drugs is not considered. To explore the added value of chemical similarity inclusion we applied proteochemometric models, combining chemical and protein target properties in a single bioactivity model. Our dataset was a large scale clinical database of genotypic and phenotypic information (in total ca. 300,000 drug-mutant bioactivity data points, 4 (NNRTI), 8 (NRTI) or 9 (PI) drugs, and 10,700 (NNRTI) 10,500 (NRTI) or 27,000 (PI) mutants). Our models achieved a prediction error below 0.5 log units. Moreover, when directly compared with previously published sequence data derived models PCM performed better in resistance classification and prediction of Log Fold Change (0.76 log units versus 0.91). Furthermore, we were able to successfully confirm both known and identify previously unpublished, resistance-conferring mutations of HIV Reverse Transcriptase (e.g. K102Y, T216M) and HIV Protease (e.g. Q18N, N88G) from our dataset. Finally, we applied our models prospectively to the public HIV resistance database from Stanford University obtaining a correct resistance prediction rate of 84% on the full set (compared to 80% in previous work on a high quality subset). We conclude that proteochemometric models are able to accurately predict the phenotypic resistance based on genotypic data even for novel mutants and mixtures. Furthermore, we add an applicability domain to the prediction, informing the user about the reliability of predictions.

6.2 Introduction

The Human Immunodeficiency Virus (HIV) was discovered and isolated as the cause of 'Acquired Immuno Deficiency Syndrome' (AIDS) in 1983.[1, 2] Over the following three decades HIV has turned into a global epidemic, the number of people living with HIV in 2010 being estimated at 34 million according to the World Health Organization.[3] Furthermore the number of people newly infected was approximately 2.7 million and 1.8 million HIV related deaths were reported,[3] hence illustrating that HIV represents one of the major illnesses of mankind today.

Infection with HIV can be contained, however not cured, by Highly Active Anti-Retroviral Therapy (HAART), which relies on a combination of three or more inhibitors from different drug classes.[4, 5] Currently more than 20 approved HIV inhibiting dugs are approved,[6] with the largest classes of drugs being formed by Protease Inhibitors (PIs), Nucleoside/Nucleotide Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs). However, while a large number of drugs is accessible to the physician (thus rendering HIV in some sense a disease that is currently 'under control' regarding the treatment options available), the question of *which drugs to use for which patient* is an exercise where more guidance would also in the current situation be of tremendous practical relevance.

6.2.1 Genetic variability. The process of replication by HIV is extremely error prone and therefore mutations in the viral genome occur frequently.[7, 8] It is these mutations that can be the basis for HIV resistance against therapy,[6] even single point mutations can cause insensitivity of HIV to treatment with all members from an entire drug class (e.g. K101P in the case of NNRTIs).[6, 9] Occurrence of these resistance conferring mutations can be contained or minimized by the nature of HAART therapy due to the combination of multiple drugs classes.[5]

However, the occurrence of high impact mutations can cause treatment failure in HAART for certain specific drug regimens. It is therefore crucial that the drug regimen is tailored to the specific viral genotype.[10, 11]

6.2.2 Personalized medicine. What is required for a tailored drug regimen is knowledge of the effect of individual mutations on the efficacy of different drugs. A rough distinction can be made between assay based methods and computational methods, with assay based methods being available since the year 1998.[12-14] Conversely, various computational methods have become available over the last decade.[15-20] Personalized prediction has been shown to perform equal to standard of care in treatment naïve patients but significantly (P = 0.02) better in patients experiencing drug failure.[17] Furthermore, computational approaches have been shown to perform equal to phenotypic assays.[21] Several methods that have been published previously, both assay-based and computational approaches, will be outlined briefly in the following.

6.2.3 Phenotypic Assays. Phenotypic assays measure the replication of HIV *in vitro* subsequent to genotype determination. Three common different phenotypic assays include: Antivirogram (AVG) by Virco (1998),[12] an assay by Walter *et al.* by the Universities of Erlangen-Nürnberg and Leuven (1999),[14] and Phenosense by Monogram Biosciences (2000).[13] Diverse readouts are employed in these assays: spectrophotometrical determination of diphenyltetrazolium bromide reduction (AVG), luminescence produced by secreted alkaline phosphatase (Walter *et al.*),[14] and luminescence by luciferase produced in the cell upon completion of one round of virus replication (Phenosense). All readouts respond in a dose dependent manner. Antiretroviral drug susceptibility is expressed as the base 10 logarithm of a numerical fold change (Log FC).

Log FC is determined by dividing the IC_{50} for inhibition of the mutated virus by the IC_{50} for inhibition of a determined wild type virus (wt). Hence, a Log FC value of 1 for a given drug – mutant pair means that the drug IC_{50} value for that particular mutant is 10 times that of the IC_{50} value for the same drug on wt. Likewise, a Log FC value of 3 for a given drug – mutant pair represents an IC_{50} value 1,000 times higher. The sequences that are defined to be wt are the HXB2 strain (Uniprot accession P04585) for AVG, [22, 23] or a recombinant pNL4-3 strain (Genbank entry M19921) for Walter *et al.* and Phenosense.[24]

6.2.4 Virtual Phenotype Approaches. From the data generated by the phenotypic assays, computational models have been produced that predict a virtual phenotype from a given genotype. Based on the large amount of Log FC data generated by AVG, Virco introduced their first computational prediction tool, Virtual Phenotype in 2000 superseded by VircoTYPE HIV-1 in 2004.[25]

VircoTYPE creates linear regression models based on the presence of mutations and pairs of mutations. Each mutation and mutation pair is given a weight factor in model training based on measured data (6,000 to 40,000 samples per drug). The sum of all weight factors for relevant mutations present in a mutant combined with the wild type weight factor then provides the predicted log FC. In a randomized clinical trial, VircoTYPE HIV-1 has been shown to perform slightly better than conventional phenotypic assays in decreasing HIV RNA concentration over a follow up period of 48 weeks (39 % of the phenotypic assay group reached HIV RNA below 400 copies/ml compared to 51 % of the VircoTYPE HIV-1 group). [21]

Next to VircoType HIV-1, another implementation of a virtual phenotype has been developed at the Max Planck Institute, called Geno2Pheno.[20] This tool has been trained on smaller data set compared to VircoTYPE. However, it has been retrospectively validated on the Stanford HIV Drug Resistance Database (Stanford Set) in 2009. [19] In this study Geno2Pheno outperformed state-of-the-art-expert based systems by finding 16.2 – 19.8 % more successful regimens.

Nevertheless, what the computational methods described here have in common is that they are solely trained on the mutation patterns and the effect these patterns have on a *single* drug.[26-28] Therefore a separate model is created for every drug. Similarity between individual amino acids is not considered (how similar are two amino acids to each other and hence how big is the impact of a mutation). Furthermore, the chemical similarity between compounds is not considered in the models. Both types of similarity information have the potential to lead to better models and prompted us to apply 'proteochemometric models', described in the following, to improve upon the current situation.

6.2.5 Proteochemometric modeling. Given that previous models did not take into account chemical information, the individual models mentioned above fail to acknowledge the chemical similarity between drugs that belong to a single class, thereby discarding very valuable information. This is the case because molecular similarity has been shown to have great predictive power when it comes to identifying which kind of *related* structures could also show activity against a given target.[29] Hence it is likely that also for established drugs, chemical similarity can improve models by explicitly taking the concept of drug – target interaction into account, which is then combined with mutational information of the drug target itself. This technique is called proteochemometric (PCM) modeling. The authors have previously reviewed the technique and it has already been successfully applied to NNRTI inhibitors of HIV Reverse Transcriptase before.[30-33]

Yet, the most important difference between this previous work and the current study is the *scale* of the mutant database used to train the models on. Previous work focused on a total of 4,792 data points,[30] 386 data points,[34] 654[31] data points, 4,495 data points,[35] or 4,024 data points,[33] whereas here a total of 288,138 data points are used. Hence, we expect a more generally applicable model resulting from the current study. Furthermore, previous work included a larger number of compounds (451 compounds) on the chemical side, and their biological activity on a total of 14 mutants. Therefore, these models described a relatively large chemical space compared to the target space, while in the current work we have reversed this situation and the models now describe a relatively large target space compared (approximately 37,000 mutants) to the chemical space (21 drugs). In addition, what is lacking in previously published PCM approaches is the power to extrapolate, thereby able to also produce a reliable prediction for novel (unknown) mutants while including a reliability measure for these predictions. These are the points we are addressing in the current work.

6.2.6 Aim of the project. In the current project it is our hypothesis that we can train a single PCM model for each of the following major HIV drug classes using the AVG data: Protease Inhibitors (PIs), Nucleoside/Nucleotide Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs). As no PCM model has ever been trained on such a large data set (the current data set is 60 times larger than the largest published HIV PCM model) our hypothesis was on the one hand to arrive at better model performance, and on the other hand to unravel more reliable rules such as the influence of point mutations on compound activity. Scientifically interesting is also the reversal in the ratio between chemical space and target space in the model training set described above.

Given the wealth of training data present, the resulting bioactivity models can be used to predict the activity of clinical ARV drugs on mutants *not present* (untested) in the data set (corresponding to a patient with a new, previously unseen genotype that needs to be treated in the clinic). For this purpose, an additional 7,798 data points have been used as a *prospective validation set*, in order to gauge predictive performance of the model in a real-world situation. These data points have been retrieved from the Stanford University database after model training and validation was completed.

6.3 Results and Discussion

6.3.1 Model Validation (Internal). The PCM modeling technique was validated in three ways. Wes started by creating a learning curve for each drug class. Learning curves plot the quality of models that are created on an increasing fraction of the data. Concurrently these models are validated on the remainder (and hence decreasing) part of the data set (supporting **Figure S1**). When given enough measures of model reliability in respect to the training set size, an estimate can be made of the optimal performance possible on said data set.

We found that all models should be able to reach a root mean square error (RMSE) < 0.5 units Log FC (see section **4.5.4** and supporting **Figure S1**), which was subsequently confirmed in the external validation which was performed per drug rather than per drug class below.

6.3.2 Model Validation (External). Models were generated on 70 % of the data set as the learning curves showed this to be the optimal split size to get a reliable performance estimate for these models. While these 70 % models give an estimate of the ability of the models to perform future predictions successfully, other additional forms of validation should also be included as we will show later on.[36] The RMSE for sequences that were present in the training set, however not in combination with the same drug, was 0.27 (PIs, **Figure 6.1C**), 0.31 (NRTIs**, Figure 6.1B**) and 0.45 (NNRTIs, **Figure 6.1A**), with an R_0^2 0.89 (PIs, **Figure 6.1C**), 0.79), NNRTIs, **Figure 6.1A**) and 0.75 (NRTIs, **Figure 6.1B**). Hence, we found that PCM was overall able to extrapolate the Log FC values for individual pairs of drug and mutant not encountered in the training set with a reliability that is comparable to the assay reliability on the current dataset (approximately 0.5 log units), with some difference encountered between the drug classes.

Hence, PCM is on this dataset able to extrapolate to novel drug-mutant pairs when the drug and mutant in question are only present in the training set individually, and not in the combination, as shown in the test set (internal validation). For sequences not present in the training set (representing predictions for previously unseen patients, or genotypes) the RMSE obtained by the model was 0.43 (PIs, **Figure 6.1F**), 0.49 (NNRTIs, **Figure 6.1D**) and 0.52 (NRTIs, **Figure 6.1E**) with an R_0^2 of 0.74 (NNRTIs, **Figure 6.1D**), 0.71 (PIs, **Figure 6.1F**) and 0.33 (NRTIs, **Figure 6.1E**), respectively. Hence, PCM is on the current dataset also able to extrapolate the Log FC values for individual pairs of drug and mutant not encountered in the training set with reliability comparable to assay reliability when the mutant in question is not present in the training set (External validation, for validation plots per individual drug please see **Figures S2-S4**).

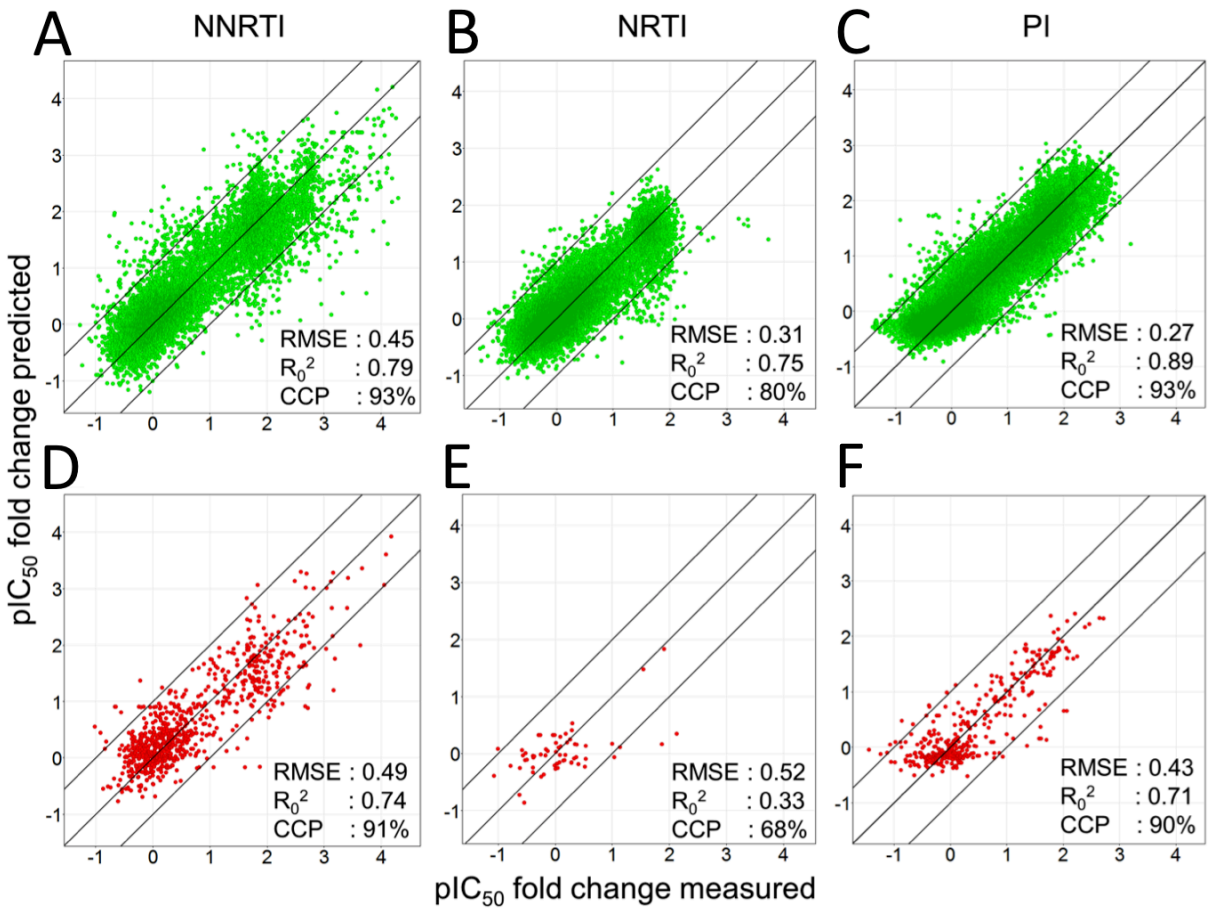


Figure 6.1: Model internal validation. (A,B,C) Our models perform robustly in both internal validation (unknown combinations of known drugs and known mutants) and (D,E,F) external validation (unknown combinations of drugs and mutants, one of which is unknown). The PIs perform the best (RMSE 0.27 log units, CCP 93% internal and 0.43 log units, CCP 90% external), followed by the NNRTIs (RMSE 0.45 log units, and CCP 93% internal and 0.49 log units, CCP 91% external) and then the NRTIs (RMSE 0.31 log units, CCP 80% internal and 0.52 log units, CCP 68% external). The range of Log FC values present in the data set is the largest for the NNRTIs, followed by the PIs and then the NRTIs.

The added value of PCM over sequence only models was also investigated (**Figures S6-S8**) in order to ensure that including chemical (ligand) information indeed improves model performance. Indeed, we found that PCM outperforms sequence only models in all drug classes. This improvement is significant for the NRTIs when performing a paired t-test (RMSE, $P < 0.01$; R_0^2, $P < 0.01$) and PIs (RMSE, $P < 0.01$; R_0^2 $P < 0.05$). The difference was not significant for the NNRTIs, while PCM did outperform sequence only models (RMSE, $P = 0.33$; R_0^2, $P=0.14$). We think this is mainly due to the large chemical diversity of the NNRTI drug class, which are similar in pharmacophoric properties but display a diverse collection of scaffolds. Since we use two dimensional chemical descriptors rather than three dimensional, PCM cannot reach the large performance difference shown for PI and NRTI. This is supported by the fact that the chemically most different NNRTI, ETR, is the only one that has a lower performance in PCM models (similarity on average 0.35, **Table S3**). Yet, the combination of the bioactivity space for individual NNRTIs is successful as NNRTIs are known to be sensitive to cross resistance, this is captured by PCM.

6.3.3 Model Validation (Clinical Cut-offs). In order to investigate clinical relevance of our work, we next incorporated the actual clinical cut-off (CCO) values. These values describe the expected response of a patient to treatment with a certain drug based on the HIV genotype. The used clinical CCO values are given in **Table S5-S7**. When we apply the CCOs to our model predictions, our models achieve an overall correctly classified percentage (CCP) of 96 % for the inhibition of mutant sequences present by a drug not present for that sequence in the data set (**Figure 6.1**).

For the sequences not present in the training set, 91 % was predicted correctly (**Table S8** and **S9**). More specifically per class, the PI scored the best (94 % correct for internal validation and 90 % correct for external validation), followed by NNRTIs (93 % correct for internal validation and 91 % correct for external validation), and lastly the NRTIs (80 % correct for internal validation and 68 % for external validation). However, it should be noted that for the NRTIs only a small number of sequences was available as validation, and only all were not very resistant, possibly leading to a biased validation. We can conclude that even prediction on sequences not present in the training set was possible, albeit slightly less than the internal validation (RMSE 0.34 log units when the sequence is known versus 0.48 when it is not). To further find the limitations of this extrapolation we employed leave-one-sequence-out (LOSO) validation.

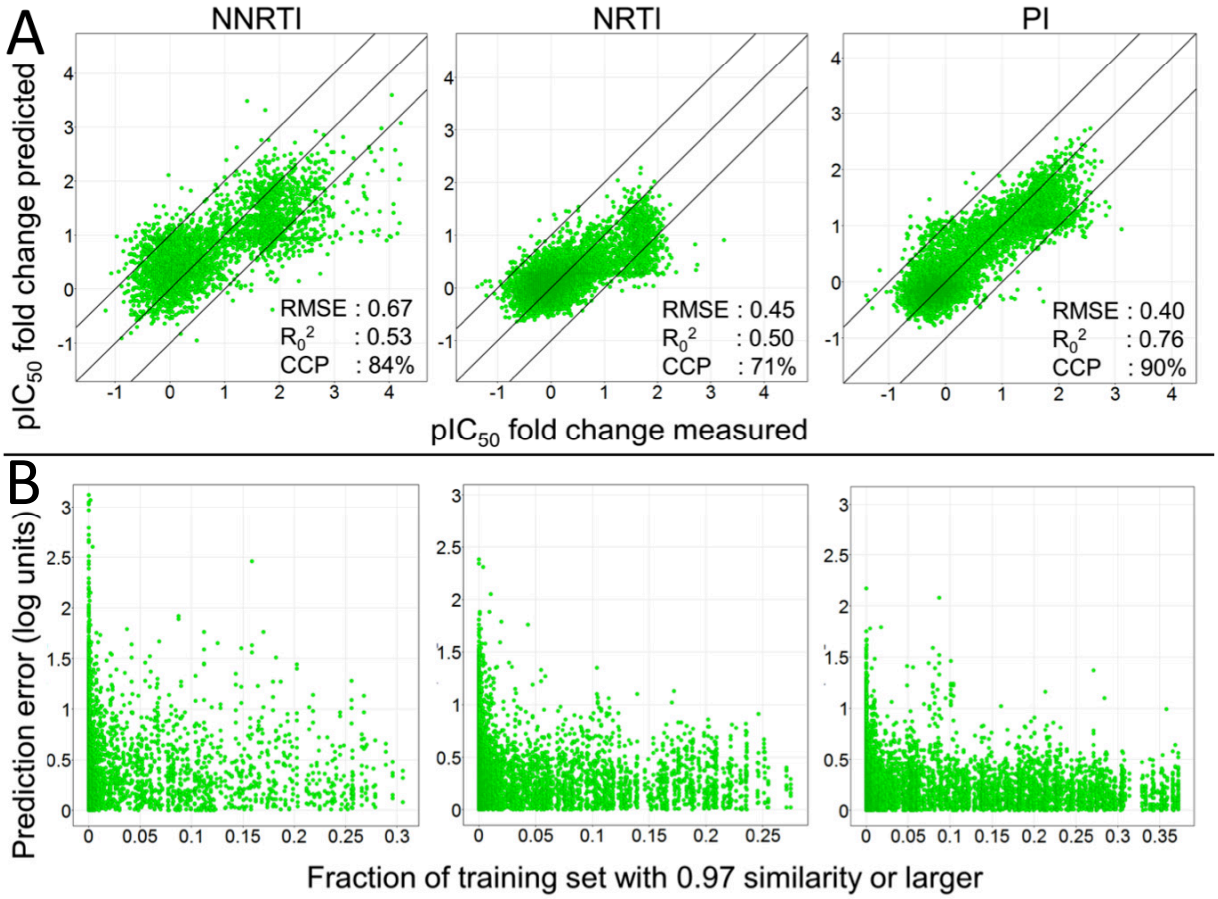


Figure 6.2: The model performance in the LOSO experiments. (A) The figure visualizes the measured Log FC for a mutant – drug pair on the x-axis. The y-axis shows the Log FC predicted for that mutant – drug pair by a model that was trained without that particular pair. Again the PIs perform the best (RMSE 0.40 log units, R_0^2 0.76, and CCP 90 %) followed by the NNRTIs (RMSE 0.67 log units, R_0^2 0.53 and CCP 84 %) and then the NRTIs (RMSE 0.45 log units, R_0^2 0.50 and CCP 74 %). (B) The density to the training set as a measure of applicability domain provides a useful estimate to predict model reliability. The x-axis shows fraction of the training set that has a similarity of 0.97 or higher to a specific mutant – drug pair. If this fraction is larger, then the prediction error (y-axis) for that pair becomes smaller as the model is better able to extrapolate from the training set. Since this fraction can be calculated before any model prediction is made, a maximally allowed prediction error can be predetermined before any model predictions are made.

6.3.4 Leave-One-Sequence-Out Validation (LOSO). LOSO validation is unique to proteochemometric approaches, since it enables the prediction of compound activities for *entirely novel genotypes* (or patients), hence estimating which treatment would be most likely to succeed in a given treatment situation. For computational reasons, our approach used a subset of approximately 1,000 mutants from the full set (4% (PR) and 9% (RT) of the total data set, respectively). Each of these sequences was left out, and a model was trained on the remaining sequences; results are shown in **Figure 6.2**.

Again, the PCM technique overall provides rather robust in modeling the current data set. Best performance can be observed for the PI model (with an RMSE of 0.40 log units, R_0^2 of 0.76 and CCP 90%), followed by the NNRTIs (RMSE of 0.67 log units, R_0^2 of 0.53 and CCP 84%) and the NRTIs (RMSE of 0.45 log units, R_0^2 of 0.50 and CCP 71%). The finding that PIs and NRTIs are easier to model than NNRTIs is in line with our finding above. What should be noted is that the NNRTI model tends to slightly underpredict the Log FC values that have been measured with a Log FC above 3.0. While those values are correctly predicted to be above 1.0 (which is an important prediction to have by itself in practice), the numerical correlation between predicted and experimental values leads to a slight, but consistent under prediction of activities in this value range.

Crucial for the application of computational models is an estimate in which cases the model can be trusted, and where it is likely to fail. In this spirit, the 'Applicability Domain' of computational models has become an important topic recently;[37] however, so far it was mainly applied to the chemical domain. This concept was extended in the current work, given the nature of PCM models, also to the protein target or biological domain where special considerations need to be taken into account. Since we are dealing with a large set of viral mutants we are unable to define a single similarity to a WT to get an idea of the applicability domain. Therefore, we chose to define the applicability domain based not only on the distance to the training set, but also on the density of neighbors in the training set (See Methods section for details). At a similarity threshold of 97 % each sequence is hence assigned a density score between 0 and 1 (0 corresponding to no sequences with a similarity of at least 97 %, and 1 corresponding to all sequences in the dataset having more than 97 % similarity to the sequence under consideration).

Figure 6.2 visualizes the 'Neighborhood Behavior';[38] if the fraction of sequences having this similarity of 97 % (X-axis) is larger (closer to 1), the maximal encountered prediction error (RMSE, y-axis) is lower (closer to 0 log units). This means that if the model can extrapolate from a larger number of sequences having a similarity of 97 % or higher, the predictions become more reliable. Performance of a practically useful model would require the largest error to be below 1 log unit; hence, given this requirement, the density of sequences in the training set should be larger than 0.15 (for PIs and NRTIs) and larger than 0.25 (for NNRTIs), respectively. Due to this numerical quantification of the 'Applicability Domain' of the model, in *biological space*, we are now able to judge in which situations the model *will be* applicable (*i.e.* is likely to generate reliable results), and in which situations it *is not* which is of crucial importance in order to gain trust into computational models.

6.3.5 LOSO Validation (Clinical Cut-offs). Further exploring the clinical relevance of this work, the CCO's were again applied to model predictions also in the case of the LOSO experiments (**Figure 6.2**). Overall the model reached a CCP of 81 % of the individual mutant – drug pairs. Moreover, 12 % of the total predictions were overpredicted, and only 7 % underpredicted. Hence our models perform robust also on sequences *that are entirely novel to the model* (**Table S10**). For the individual classes, the image is very similar to that in the external validation, the PIs perform the best (90 % correct), followed by the NNRTIs (84 % correct) and lastly the NRTIs (71 % correct).

In the text above we have thoroughly validated our models and they have shown to be robust in modeling HIV resistance to PIs, NNRTIs and NRTIs. This was confirmed for known sequences in an unknown combination with a drug but also for unknown sequences in an unknown combination with a drug. Hence we conclude that our models describe the drug – target interaction space, therefore it is very interesting to investigate how our models actually derive these Log FC values from the contributions individual mutations make.

6.3.6 PCM compared to sequence only models. To compare the performance of our PCM models with state of the art models trained on sequence data only, we used a data set previously published by Van der Borght *et al.*[39] We explicitly selected for each class the 150 sequences that were predicted most inaccurate, representing the most difficult sequences to predict (representing mutants that seem to exhibit a different resistance profile). Moreover, most of these sequences contained mixtures (several mutations present on a single position) that had been discarded from our PCM training set. The purpose of this validation was therefore twofold, to assess the performance of PCM when compared to sequence only models, and secondly to assess if the PCM models can deconvolute the effect of individual mutations to make accurate predictions for mixture sequences.

The results of this validation are shown in **Figure 6.3** and **Table 6.1**. Our PCM models clearly outperform sequence only models. For each class the PCM models predict the Log FC more accurately. This is indicated by the smaller RMSE (0.53 log units versus 0.68 log units for the NRTIs; 0.65 log units versus 0.75 log units for the PIs, and 1.1 log units versus 1.3 log units for the NNRTIs) and also by a higher CCP (68 % versus 54 % for the NRTIs, 78 % versus 75 % for the PIs, and 89 % versus 78 % for the NNRTIs). For several PIs, the sequence only models perform marginally better when measuring by the correlation coefficient; however as these values are systematically slightly overpredicted, the prediction error is still larger than for the PCM models.

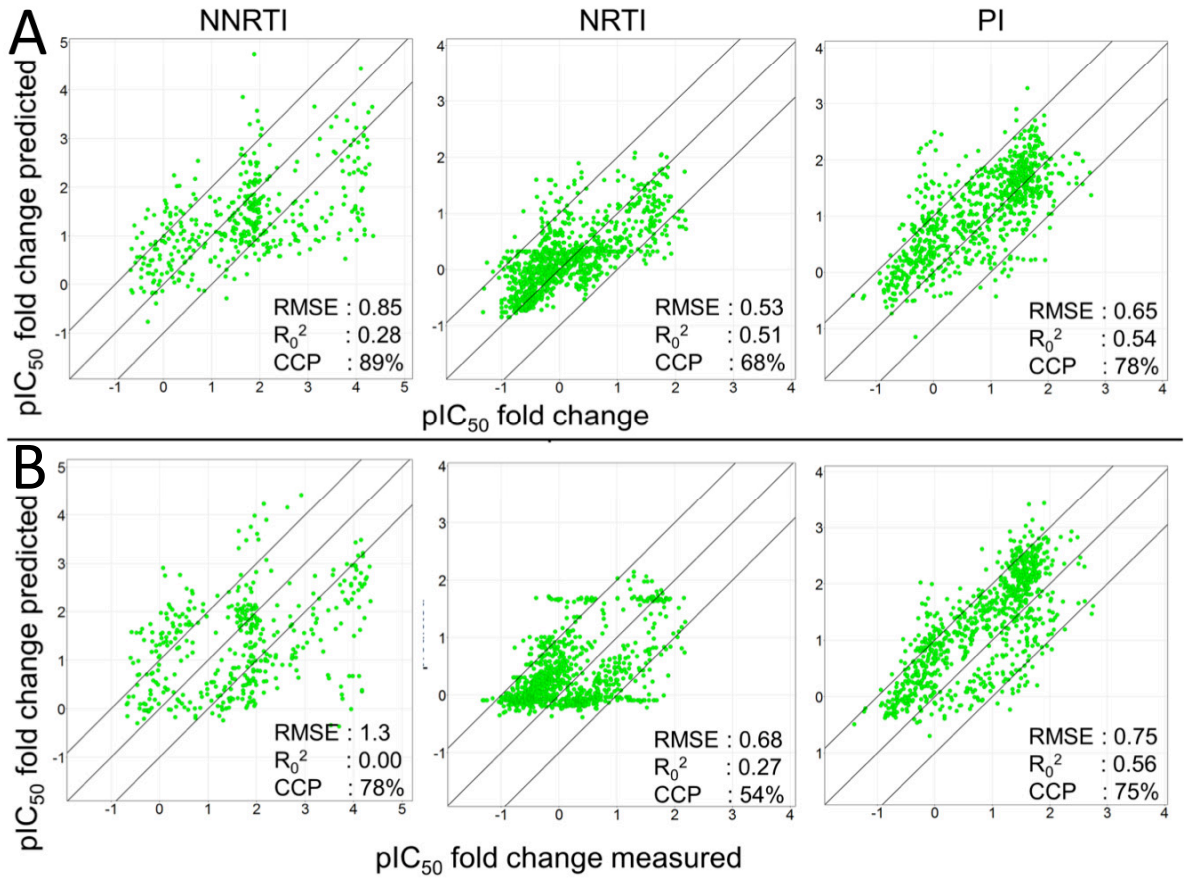


Figure 6.3: Performance of PCM based models compared with sequence based models for the 150 most difficult sequences as published by Van der Borght e*t al.* The PCM models (A) perform better as they have a lower prediction error for each drug class (0.53 log units versus 0.68 log units for the NRTIs; 0.65 log units versus 0.75 log units for the PIs, and 1.1 log units versus 1.3 log units for the NNRTIs) than the sequence based models (B). Clearly the NNRTIs are most difficult to predict. Note that these sequences contain a large fraction of mixture sequences, which were not present in the PCM training set but were present in the sequence only training set. In addition, the PCM models also reach a higher CCP compared to the sequence only models.

Furthermore, when we limit ourselves to only predicting the Log FC for mutant mixtures, PCM still outperforms sequence only models (**Table S10** and supporting **Figure S12**). This is even the case while our PCM models were trained without mixture sequences in the training set whereas these were present in the training for the sequence only models. A large fraction of these mixtures sequences show a low value for the 97 % similarity density, hence we would expect the models to perform suboptimal on these sequences. The applicability domain measure therefore also holds in this case. These results underline the added value of PCM models over sequence only models, hence we wanted to test the performance of our models prospectively on a clinical data set to judge their clinical relevance.

Table 6.1: Performance of PCM compared to sequence only models

RMSE (Log units)	R_0^2	RMSE Sequence only (Log Units)	R_0^2 Sequence only	Grouping
0.66 (±25)	0.41 (± 0.19)	0.80 (± 0.29)	0.22 (± 0.41)	Drug (average)
0.65	*0.54*	*0.75*	*0.56*	*PI (Class)*
0.59	0.64	0.67	0.66	APV
0.67	0.57	0.83	0.50	ATV
0.80	0.39	0.79	0.49	DRV
0.62	0.54	0.76	0.59	IDV
0.65	0.60	0.83	0.67	LPV
0.63	0.49	0.73	0.48	NFV
0.63	0.52	0.76	0.57	SQV
0.53	0.41	0.55	0.42	TPV
0.85	*0.28*	*1.3*	*0.00*	*NNRTI (Class)*
0.93	0.39	1.1	0.10	ETR
1.5	0.12	1.8	0.00	EFV
0.72	0.00	0.95	0.00	NVP
0.53	*0.51*	*0.68*	*0.27*	*NRTI (Class)*
0.67	0.49	0.83	0.31	3TC
0.41	0.46	0.53	0.15	ABC
0.59	0.45	0.75	0.20	AZT
0.45	0.27	0.54	0.00	D4T
0.42	0.35	0.51	0.10	DDI
0.65	0.51	0.90	0.20	FTC
0.43	0.36	0.59	0.00	TDF
0.66	***0.42***	***0.80***	***0.30***	***Overall***

Validation parameters were calculated using different forms of grouping to give an unbiased error estimate. The table shows that our PCM models perform better than sequence only models. This is indicated by the regression validation parameters RMSE and R_0^2. While it should be noted that for some of the PIs, the sequence only models tend to have a slightly higher R_0^2, they also have a much higher RMSE.

6.3.7 Model Interpretation (Known Resistance Mutations). The aim of this feature importance investigation was to explain the *average* reduction in drug affinity that the presence of an individual mutation causes. Firstly, we investigated the effect of several known mutations from literature. To this end we compared the features selected as being significant by our model to the mutational overviews published by Johnson *et al.*[6, 40]

Figure 6.4 shows the impact of selected mutations on NNRTI affinity. Overall, while there is a significant amount of cross-resistance, each of the NNRTIs still possesses its own distinct resistance profile, in agreement with the importance of personalized HIV treatment approaches.

Furthermore, the impact of individual mutations, shown as a darker shade of red, varies per drug and is in line with literature data.[6, 18] (For an explanation of the abbreviations see supporting **Table S1**) For instance, mutation K103N has a rather specific pattern as it confers resistance to Nevirapine, Efavirenz, and Delavirdine but not to Etravirine.[6, 18] This pattern is reproduced by our model. Furthermore, V179F is known to lead to Etravirine resistance but to have less effect on Nevirapine, Efavirenz, and Delavirdine,[6, 18] a resistance profile that can also be reproduced based on our dataset. Some mutations are slightly underestimated, these include V90I and V106I. Another interesting observation is that mutations Y188C and G190A are predicted to render HIV *more sensitive* to Etravirine according to our model. This finding is in agreement with work by Vingerhoets *et al.* [41]

Related analyses for NRTI resistance and PI resistance have been included in the supporting information (supporting **Figure S8** and supporting **Figure S9**). Specific NRTI mutations that were accurately reproduced include K65R, Q151M, and T215Y, while mutations M41L and M184V are slightly underestimated, compared to previous studies.[6] For the PIs mutations that are accurately reproduced include D30N, I50L, V82S, and I84, while the I64L and I93M mutations are assigned less importance than in previous work.[6]

Hence, the PCM models applied in this study are able to reproduce known resistance patterns as outlined above. This led us to the next step of the study, the identification of *novel* mutations (present in our data set but not previously published) which are found to confer cross resistance to antiretroviral treatments. This work is similar to previous work by Van der Borght *et al.*[39] but here we focus on both cross resistance conferring mutations *and* drug specific mutations. Furthermore we apply the method to all three major classes of anti-HIV drugs rather than one and can do so directly from our models.

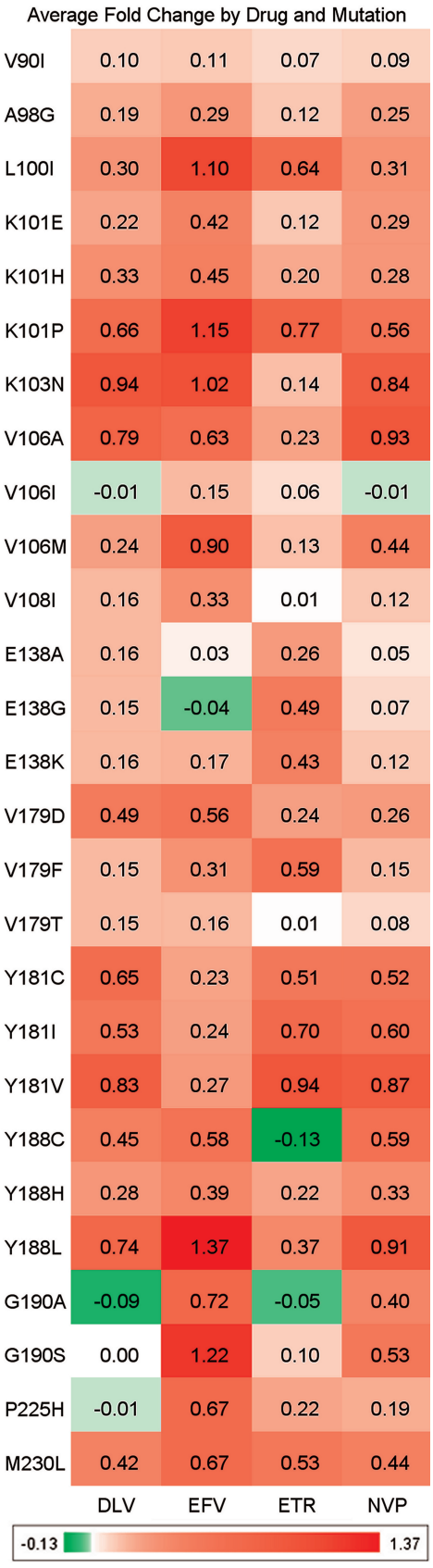


Figure 6.4: Model interpretation, known mutations that lead to NNRTI (cross) resistance. The pattern produced by our model correlates with literature. [6, 18] In particular the specific profiles of V106I, Y188C and G190A are reproduced well. Values in the cells represent Log FC.

6.3.8 Model Interpretation (Cross Resistance-Conferring Mutations). To identify cross-resistance as part of the current study, we were limiting ourselves to mutations that have a *negative* effect on the majority of drugs in a single class. However, in case of particular interest in the resistance profile of a particular drug this analysis can also be performed on the individual-drug level subsequently. We selected mutants based on the following conditions: occurrence in the data set more than once; average Log FC for all compounds above 0.4; standard deviation over this average below 0.4. Known mutations as published in literature were discarded.[6, 40, 42, 43] With these filters a number of novel resistance conferring mutations could successfully be identified which are listed in **Table 6.2** to **Table 6.4** (For an explanation of the abbreviations see supporting **Table S1**). Mutations identified have a high impact on drug affinity and which lend themselves to experimental validation, for instance in the case of NNRTI and NRTI resistance conferring mutation T216M. The full set of individual mutations (both known and novel) and their effect is included in the supporting information as delimited text files.

Table 6.2: Novel resistance conferring mutations derived from the dataset (NNRTI).

Mutation	DLV	EFV	ETR	NVP	Average Log FC
P9T	0.36	1.01	0.65	0.46	0.62
E79D	0.34	0.55	0.61	0.34	0.46
K101S	0.38	0.73	0.31	0.44	0.47
K102Y	0.72	0.53	0.47	0.77	0.62
S156A	0.8	1.2	0.76	0.67	0.86
M164L	0.26	0.89	0.51	0.62	0.57
T216M	0.97	1.47	0.01	0.84	0.82
Y232H	0.47	0.62	0.36	0.47	0.48
R307M	0.92	0.14	0.28	0.35	0.42
Average Susceptibility	0.58	0.79	0.44	0.55	-

The value in the different drug columns indicates the average Log FC in the presence of this mutation. While these mutations have been selected to confer some resistance to all NNRTIs, each drug still has a distinct profile. Efavirenz is the most sensitive (average Log FC 0.79) and Etravirine the least (average Log FC 0.44) with Nevirapine (average Log FC 0.55) and Delavirdine (average Log FC 0.58) in between.

Table 6.3: Novel resistance conferring mutations derived from the dataset (NRTI).

Mutation	3TC	ABC	AZT	D4T	DDC	DDI	TDF	FTC	Average Log FC
I63V*	0.22	**n/a**	1.07	0.53	**n/a**	0.52	0.01	0.36	0.45
I202M*	0.23	**n/a**	0.73	0.68	0.51	0.57	0.45	0.39	0.51
R206M	0.90	0.42	0.54	0.01	0.15	0.17	0.24	0.92	0.42
T216M	0.88	0.51	0.66	0.12	0.20	0.27	0.38	0.94	0.50
E298K*	0.33	0.43	0.44	0.38	0.65	0.32	**n/a**	**n/a**	0.43
Average Susceptibility	0.51	0.45	0.69	0.34	0.38	0.37	0.27	0.65	-

The value in the different drug columns indicates the average Log FC in the presence of this mutation, when not available in the data set the value is denoted 'n/a'. Mutations indicated with an asterisk were incompletely tested on all drugs in the data set. Like the NNRTI resistance mutations, each mutation displays a different resistance profile over all drugs. AZT is seen to be the most susceptible (average Log FC 0.69) and TDF the least susceptible (average Log FC 0.27).

Table 6.4: Novel resistance conferring mutations derived from the dataset (PI).

Mutation	APV	ATV	DRV	IDV	LPV	NFV	RTV	SQV	TPV	Average Log FC
Q18N	0.55	0.52	0.56	0.61	0.58	0.49	0.56	0.50	0.65	0.56
V32T*	0.63	0.65	0.07	0.67	0.45	0.67	0.68	0.81	**n/a**	0.58
N88G	0.39	0.97	-0.27	0.77	0.18	1.14	0.10	0.49	0.22	0.44
Average Susceptibility	0.52	0.71	0.12	0.68	0.40	0.77	0.45	0.60	0.44	-

The value in the different drug columns indicates the average Log FC in the presence of this mutation, when not available in the data set the value is denoted 'n/a'. Mutations indicated with an asterisk were incompletely tested on all drugs in the data set. Here Nelfinavir is the most susceptible (average Log FC 0.77) and Darunavir the least (average Log FC 0.12).

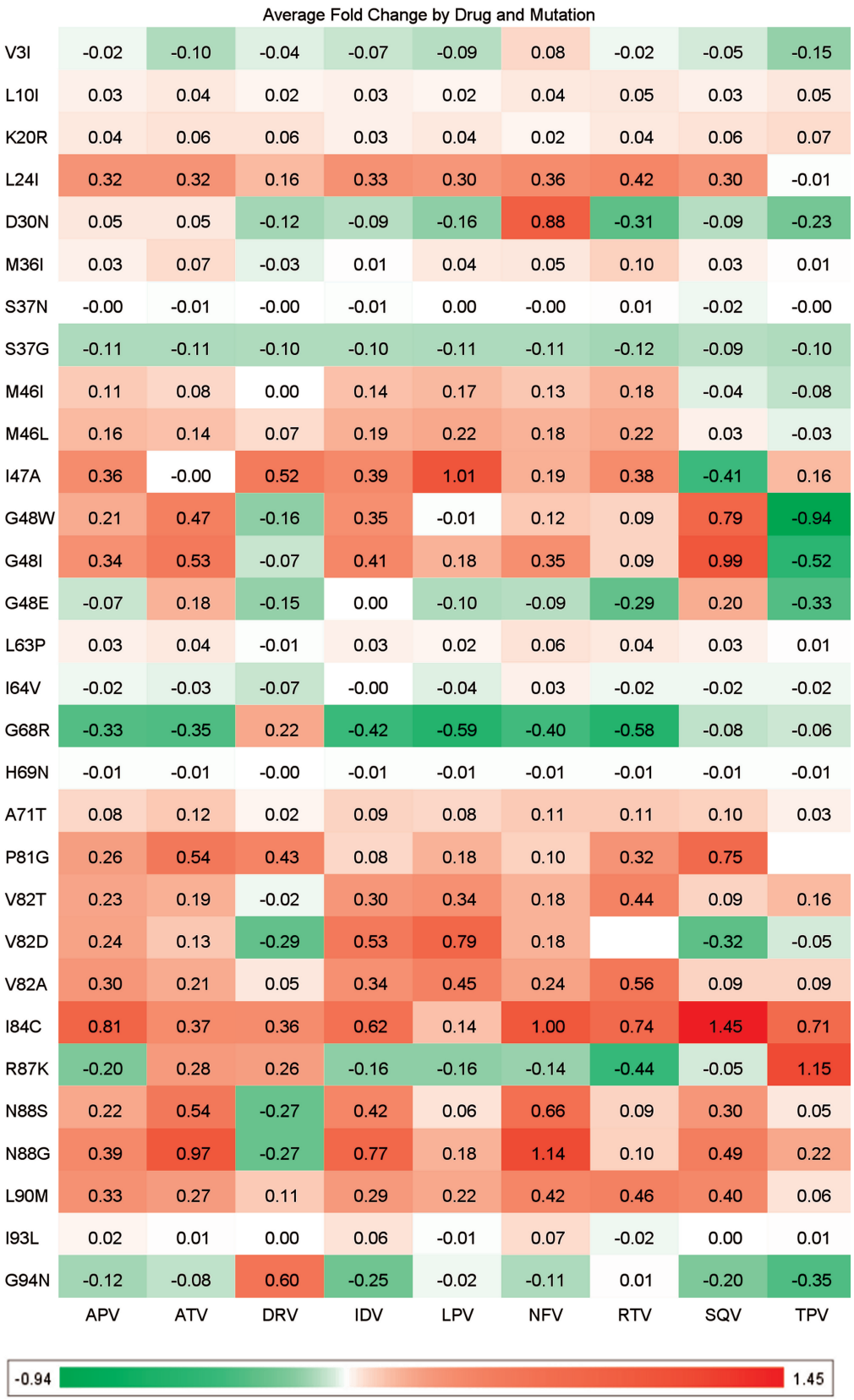


Figure 6.5: Model interpretation, mutations leading to drug specific resistance. Shown are the 30 mutations that have the most diverse effect over the different members of the PI drug class. The figure contains a number of known mutations (e.g. M46L,[6] A71T, [6] V82A,[6] V82S[6]) but also several novel mutations (e.g. G48W, N88G). Values in the cells represent Log FC.

6.3.9 Model Interpretation (Drug-Specific Resistance-Conferring Mutations). We furthermore analyzed not only mutations that cause cross-resistance, but also those with a particular effect on a specific drug treatment alone. The goal here was to identify mutations that lead to large resistance for one drug but are still sensitive for another drug from the same class. Hence this knowledge can be of high importance in a clinical setting. For the PIs the 30 most interesting mutations (defined as those mutations that have the most diverse effect on the different drugs), are shown in **Figure 6.5** (while corresponding figures for the NNRTIs and NRTIs are included in supporting **Figure S10** and supporting **Figure S11**). In those figures we can observe several mutations that lead to resistance for a single drug (Log FC on average > 0.5) and at the same time lead to higher sensitivity for another drug (Log FC on average < 0.0).

For instance, the G48W mutant is sensitive to Darunavir and Tipranavir, while showing some degree of resistance to all other PIs. Furthermore, R87K is resistant to Atazanavir, Darunavir, and Tipranavir, but sensitive to all other drugs in the dataset. This could indicate that at this point the mutant has over-adapted to the host environment, including the drug, hence rendering the mutant very sensitive to changes in this environment. Finally, N88G seems to only be sensitive to Darunavir, while conferring resistance to all other PIs in the dataset. Information of this type is of high relevance to prescribe the optimal drug for an individual patient, by being able to link the viral genotype to the clinical phenotype in a real-world situation. Applying these models in a real world situation on unseen clinical data is exactly what we implemented in the following paragraphs.

6.3.10 Personalized predictions (Stanford University Data). Given a sequence of PR and RT (and hence, a viral genotype of a patient to be treated), our models are able to predict which drugs will provide the best treatment combination (corresponding to the lowest resistance to a particular drug, as measured via the lowest Log FC). To accurately estimate our model performance in personalized predictions, in the final step of this study we performed a prospective model validation. Apart from only focusing on unseen data, in order to establish agreement of our modeling procedure with other approaches, we also employed data from an entirely different source – namely, for sequences obtained from the Stanford University HIV Drug Resistance Database (Stanford Set).[18, 44]

6.3.11 Personalized predictions (Model performance). Applied to the Stanford Set, the PCM models developed in the current work show an average RMSE of 0.52 log units, with the average R_0^2 being 0.59. Compared to the models above, this is a slightly larger error compared to the validation on Virco data, which was below 0.50 log units. (It should be noted that this is very diverse data, including historical literature data of which we cannot estimate reliability.) The PI model again performs the best (with an RMSE of 0.44 log units and an R_0^2 of 0.75), while the NNRTIs are predicted with the largest error (with an RMSE of 0.68 log units and an R_0^2 of 0.65), which is the result of a number of outliers (see **Figure 6.6** and explanation below). The NRTI model exhibits the lowest correlation coefficient (R_0^2 0.39 and RMSE 0.61 log units), mostly due to the relatively small range of Log FCs present in the data set. However, also in this case we observe a correlation between the density of sequences with a 97 % similarity in the training set and modeling error, also allowing us to establish the Applicability Domain of the model throughout.

6.3.12 Personalized predictions (Discussion of Outliers). With the NRTIs and some NNRTIs there are outliers to the Applicability Domain we established, meaning that expected and observed errors exhibited some differences. (Note that this is usually the case, the Applicability Domain concept being a concept based on error distributions and likelihoods, not certainties, that a given error will be obtained in a given situation.) It was found that these outliers were obtained from only a small number of references (RefIDs) from the Stanford DB. References 369, 414, 649 all contained the M184V and T215Y mutations that are also known to differ between AVG and Phenosense. Furthermore there was a major discrepancy between the Log FC values reported for AZT on similar mutant which was > 2000 (log value 3.3) in one reference, while being reported as low as 28 (log value 1.4) from another source.[45-47] Reference 789 contained a sequence carrying a deletion at position 69, which was not taken into account by our model.[48] Reference 947 linked to unpublished data and could therefore not be verified. Finally, reference 1261 is underpredicted for both the NRTI tested sequences and NNRTI tested sequences and we could not identify an apparent cause for this behaviour.[49] (More detailed results are listed in **Table 6.5**.) The table shows that performance per drug is very good with a low RMSE (an average RMSE of 0.54 log units; with two outliers, AZT and FTC, exhibiting an RMSE of > 0.90 log units). Overall, when the results are grouped per literature reference number (which is included in the data set) the average quality decreases and the standard deviation increases, indicating that differences between reported Log FC changes in literature exist and this could adversely affect model performance.

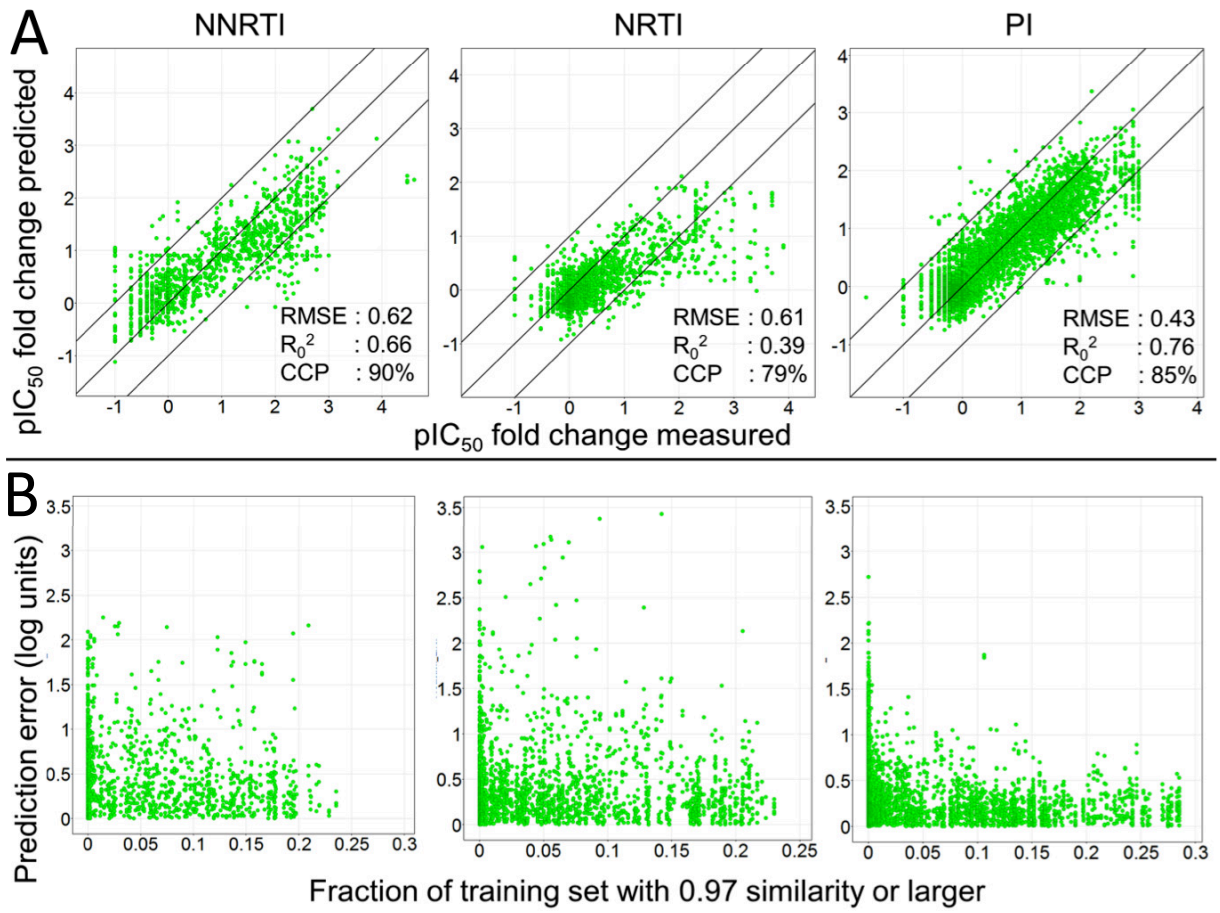


Figure 6.6: Model performance predicting the Stanford University data set. (A) The isolates predicted were not included in the training set, still performance is robust. The NNRTIs perform the best (RMSE 0.62 log units and CCP 90 %), followed by the PIs (RMSE 0.43 log units and CCP 85 %) and then the NRTIs (RMSE 0.61 and CCP 79 %). (B) The density to the training set as a measure of applicability domain provides a useful estimate to predict model reliability. The x-axis shows fraction of the training set that has a similarity of 0.97 or higher to a specific mutant – drug pair. The larger this fraction, the smaller the prediction error (y-axis) for that pair as the model is better able to extrapolate from the training set.

6.3.12 Personalized predictions (Discussion of Outliers). With the NRTIs and some NNRTIs there are outliers to the Applicability Domain we established, meaning that expected and observed errors exhibited some differences. (Note that this is usually the case, the Applicability Domain concept being a concept based on error distributions and likelihoods, not certainties, that a given error will be obtained in a given situation.) It was found that these outliers were obtained from only a small number of references (RefIDs) from the Stanford DB. References 369, 414, 649 all contained the M184V and T215Y mutations that are also known to differ between AVG and Phenosense.

Furthermore there was a major discrepancy between the Log FC values reported for AZT on similar mutant which was > 2000 (log value 3.3) in one reference, while being reported as low as 28 (log value 1.4) from another source.[45-47] Reference 789 contained a sequence carrying a deletion at position 69, which was not taken into account by our model.[48] Reference 947 linked to unpublished data and could therefore not be verified. Finally, reference 1261 is underpredicted for both the NRTI tested sequences and NNRTI tested sequences and we could not identify an apparent cause for this behaviour.[49] (More detailed results are listed in **Table 6.5**.) The table shows that performance per drug is very good with a low RMSE (an average RMSE of 0.54 log units; with two outliers, AZT and FTC, exhibiting an RMSE of > 0.90 log units). Overall, when the results are grouped per literature reference number (which is included in the data set) the average quality decreases and the standard deviation increases, indicating that differences between reported Log FC changes in literature exist and this could adversely affect model performance.

6.3.13 Personalized predictions (Clinical Cut-offs). While, each assay uses its own set of CCO values tuned for the respective assay, we used values supplied by Virco and Rhee *et al.*[44] Clinical classification is included in **Table 6.5** as 'Correctly Classified Percentage', and it represents the percentage of the data points that was classified correctly. Our model classifies the response correctly in 84 % of the cases. The average performance when grouped per individual drug class was very good (PI 85 %, NNRTI 89 % and NRTI 79 %). Also noteworthy is that the model bias is towards over prediction rather than under prediction, something that is not always mentioned in literature but is especially relevant in a clinical setting.

Previous work on a *high quality filtered subset* of our Stanford DB set reached 80 % correct predictions of phenotype from genotype on average (PI 78 %, NNRTI 83 % and NRTI 75 %).[44] Other work indicates that an expert panel reaches up to 44 % correct predictions.[50] The two outliers in the NRTI class are d4T and TDF, for which an apparent discrepancy between AVG data and Phenosense data has previously been described.[51]

Table 6.5: Personalized prediction examples for isolates not present in the original data set.

RMSE (Log units)	R_0^2	Correctly Classified Percentage	Overpredicted Percentage	Underpredicted Percentage	Grouping
0.54 (±0.28)	0.56 (±0.27)	0.85 (±0.13)	0.09 (±0.11)	0.06 (±0.09)	RefID (average)
0.45 (±0.33)	0.62 (±0.34)	0.84 (±0.24)	0.11 (±0.20)	0.06 (±0.15)	IsolateName (average)
0.44 (±0.34)	0.62 (±0.34)	0.84 (±0.24)	0.10 (±0.20)	0.06 (±0.15)	SeqID (average)
0.54 (± 0.18)	0.58 (±0.19)	0.83 (±0.10)	0.11 (±0.10)	0.06 (±0.06)	Drug (average)
0.44	*0.75*	*0.85*	*0.11*	*0.03*	*PI (Class)*
0.43	0.74	0.86	0.10	0.04	ATV
0.37	0.75	0.72	0.28	0.00	IDV
0.39	0.83	0.91	0.03	0.06	LPV
0.44	0.76	0.9	0.05	0.04	NFV
0.44	0.78	**0.91**	**0.05**	**0.03**	**RTV**
0.49	0.75	**0.88**	**0.07**	**0.05**	**SQV**
0.52	0.38	0.70	0.20	0.02	TPV
0.68	*0.65*	*0.89*	*0.05*	*0.06*	*NNRTI (Class)*
0.64	0.63	**0.83**	**0.10**	**0.07**	**DLV**
0.60	0.70	1.00	0.00	0.00	EFV
0.76	0.65	0.87	0.04	0.09	NVP
0.61	*0.39*	*0.79*	*0.12*	*0.09*	*NRTI (Class)*
0.47	0.49	0.85	0.09	0.07	ABC
0.90	0.56	0.84	0.09	0.07	AZT
0.41	0.37	0.64	0.12	0.23	D4T
0.42	0.35	**1.00**	**0.00**	**0.00**	**DDC**
0.39	0.41	0.74	0.17	0.10	DDI
1.01	0.66	0.85	0.00	0.15	FTC
0.44	0.12	0.66	0.30	0.04	TDF
0.53	***0.65***	***0.84***	***0.10***	***0.06***	***Overall***

Validation parameters were calculated using different forms of grouping to give an unbiased error estimate. Class wide values are indicated in italic and the global average performance is indicated in bold and italic. For larger groups (RefID, SeqID, Isolatename and per drug) the average value and standard deviation are given. For three drugs (RTV, DLV, DDC) no Virco cut-off was available, here the Stanford cut off was used for both, for SQV no Stanford cut-off was available so the Virco cut-off was used for both. The table shows that our PCM models perform robustly in predicting the Log FC as indicated by the regression validation parameters RMSE and R_0^2. More importantly, the correctly classified percentage is 84% overall.

6.4 Conclusions

In this work we report the construction of robust PCM models, based on 200,000 bioactivity data points measured against different HIV genotypes. In total, the model contained information on a total of 4 (NNRTI), 8 (NRTI) or 9 (PI) drugs combined with 10,700 (NNRTI) 10,500 (NRTI) or 27,000 (PI) mutants. Given the nature of the PCM modeling procedure employed in this work, we were able to combine all resistance profiles of the three above drug classes in three single models, hence focusing on very large *target space* (tens of thousands of different proteins) in this work. Both in internal and prospective validations our model showed performance comparable to assay reliability and better than sequence only models; moreover, model interpretation has been performed to identify *novel* resistance-conferring mutations that lead to resistance to *all* drugs in a class, such as T216M in the case of RT. In addition, we can use these models to find mutations that lead specific sensitivity (G48W in PR) or resistance (G68R in PR) to a single drug within a class.

Another application of our models is personalized drug regimen predictions. We have shown that our models are able to predict clinical resistance with a high degree of reliability. This reliability is formed by a 95 % CCP when predicting clinical response for Antivirogram data, which is the assay models were trained on, similar studies reached 80 % CCP when predicting values for the assays they trained on. Furthermore, the CCP and is as high as 81 % when predicting clinical response for *unknown* mutants. The novelty is formed by reliable predictions on *unknown* mutants and even *unknown mixtures*. Finally, the CCP is 84 % when predicting clinical response for clinical isolates obtained from very diverse sources (including historical literature data and data from different assays), indicating that the model is robust and predictive.

We attribute the better performance of PCM to two reasons. Firstly our models are trained on a very large co-linked dataset. This large training set not only minimizes the influence and bias caused by single experimental error, it also allows the model to detect global patterns that are consistent over both genotype (sequence similarity) and chemo type (drug similarity). The second reason is related to the first, as the encoding of the *full sequences* using physicochemical properties rather than presence or absence of mutations allows for a better similarity measure between two sequences.

6.5 Methods

6.5.1 Data Set. The main data set (**Table 6.6**) was obtained from Virco (Beerse, Belgium) and consisted of mutants (both PR and RT sequences) and fold change (Log FC) in pIC_{50} (log units) data in the AVG assay collected by Virco up to January 2011.[12, 25, 28] Mixtures, consisting of multiple mutants that were identified in a single clinical isolate) were removed from the set. The Log FC data was used as is, since it already consisted of log units difference to a single mutant defined as wild type. The wild type was defined as the HXB2 isolate (Uniprot accession P04585 and Genbank accession K03455).[22]

Table 6.6: Description of the data set used in the current study (Obtained from Virco).

Target	Amino acids	Binding Site	Drug Class	Drugs	Mutant Sequences	Data points
Reverse Transcriptase	400*	Orthosteric	NRTI	8	10,501	72,727
Reverse Transcriptase	400*	Allosteric	NNRTI	4	10,723	35,249
Protease	99	Orthosteric	PI	9	27,081	180,162

* For Reverse Transcriptase only the first 400 amino acids were sequenced. The total size of the data set is unlike any other data set used in PCM.

6.5.2 Mutant descriptors. Sequences were subsequently encoded using the first three Z-scales.[52, 53] For PR the full sequence was used and for RT only the first 400 amino acids were sequenced as the final 160 residues form an RnaseH domain and are not directly relevant in (N)NRTI resistance. These Z-scales were subsequently used to train models.

6.5.3 Drug descriptors. Structures of the drugs were normalized and ionized at pH 7.4, they were assigned 2D coordinates and subsequently converted to Scitegic circular fingerprints.[54, 55] All this was done in Pipeline Pilot Student Edition version 6.1.5.[56] Circular descriptors provide individual substructures and treat these as a feature of a compound. These substructures are centered iteratively around all atoms of the compound with a specified maximal diameter and they have been shown to give very high retrieval rates in comparative studies.[57] The following circular fingerprints were used (with the underscore denoting the maximal bond diameter): NRTIs used ECFP_10, NNRTIs used ECFP_8 and PIs used ECFP_12.

In order to create a numeric descriptor for each drug, a similarity matrix was constructed using the fingerprints and based upon the Tversky Similarity coefficient.[58] Here fingerprints were converted to a fixed length array of counts with maximal length of 256 bits, the value for α was 0.1 and the value for β was 0.9, putting more weight on the unique features of the target molecules compared to the reference molecule. For each drug in a class, the similarities to all other drugs from that class were then used as a descriptor (**Tables S2** – **S4**).

6.5.4 Machine learning. Models were constructed in the academic version of Pipeline Pilot 6.1.5 using the R-statistics package.[59] Support vector machines (SVM) as coded in the e1071 package were used for model creation.[60] Parameters gamma and cost were tuned over an exponential range and epsilon was set at 0.25. It has been shown that setting epsilon to the approximate data error is the optimal value for training.[61] The optimal model was determined using 5 fold cross validation before proceeding to external validation of the model. The parameters used for validation were R_0^2, R^2, and RMSE.[36, 62]

6.5.5 Density based Applicability Domain. As our models are trained on a database of different HIV mutants, applicability domains based on a single wild type sequence are expected to perform sub optimal. Rather we choose to determine the applicability domain based on the density of the nearest neighbors in the training set. This density was expressed as the fraction of the total number of sequences meeting a certain similarity criterion. Therefore this density score will be between 0 (0 %, no sequences meeting the similarity criterion) and 1 (100 %, all sequences meeting the similarity criterion). We calculated the density at a large number of similarity thresholds between 99 % and 70 %. Optimal performance was reached at 97 %, similarity defined as 1 minus the euclidean distance. Furthermore, this similarity was based on full sequence similarity rather than binding site similarity.

Hence for each sequence, the total number of sequences being 97 % similar or more can be between 0.0 (none) and 1.0 (all). We found that in practice the total fraction did not exceed 0.3 (30 % of the sequences in the training set 97 % similar or more).

6.5.6 Learning Curves. The learning curves provide an estimate for the maximal performance that can be achieved on these data sets, simultaneously they represent external validation. The learning curves show that the models gradually improve when trained on a larger data set. The results show that PCM is not only able to create models on this data, but also that these models are robust with good validation parameters. The PI model shows the best performance, RMSE < 0.40 log units when trained on 5 % of the full set and < 0.30 log units when trained on 70 % of the data set. The NNRTI model the worst performance, RMSE = 0.70 log units when trained on 5 % of the full set and < 0.50 log units when trained on 70 % of the data set (supporting **Figure S1**).

6.5.7 Y-Scrambling. Subsequent to learning curve creation, y-scrambled models were created. Here the measured value (*i.e.* Log FC) was randomly permutated over the drug – mutant combination. The rationale being that no correlation should remain as the presence of a certain mutation will no longer be associated with a lower Log FC value but with mixed Log FC values. Supporting **Figure S12** to supporting **Figure S14** display the lack of correlation between measured and scrambled values.

Models that were trained on this scrambled set and validated on 30 % the data that was kept unscrambled produced very high RMSE values. These values were (in log units); 0.83 (PIs, versus 0.27 for predictive models), 1.10 (NRTIs, versus 0.31 for predictive models) and 1.11 (NNRTIs, versus 0.45 for predictive models). Furthermore, the values for the R_0^2 were very low; -0.06 (PIs, versus 0.89 for predictive models), -0.20 (NRTIs, versus 0.75 for predictive models) and -0.21 (NNRTIs, versus 0.79 for predictive models) (supporting **Figure S15** to supporting **Figure S17**). Finally the cross validation parameters for the models trained on these scrambled sets demonstrated a lack of correlation; RMSE in log units was highly similar to the external validation; 0.87 (PIs), 1.11 (NRTIs) and 1.12 (NNRTIs). The corresponding correlation coefficient was 0.00 for all three models.

6.5.8 Model Interpretation. To determine the effect of individual residues, for each sequence each residue was mutated back to wild type *in silico* by replacing the descriptors of the mutant amino acid with the descriptors of the corresponding wild type residue.[33] Subsequently for all drugs the model prediction on the original mutant sequence was compared with the prediction of the model on the in silico changed mutant sequence. The difference was interpreted as the change in pIC_{50} induced by that particular residue, hence providing model interpretability. Changes that led to a 0 value shift in pIC_{50} were removed in the calculation of the average influence of mutations in a particular position, since in all cases this was caused by substitution of identical amino acids.

6.5.9 Known resistance mutations. Known resistant mutations were retrieved from earlier publications by Johnson *et al.* and compared to our model interpretation.[6, 40] While these papers only mention high impact mutations and are gathered over the full population, they are a good frame of reference for our model interpretation. We used both the most recent publication and one from 2006 as Delavirdine (DLV) has been removed from these overviews due to the fact that it is only used rarely.

6.5.10 Cross Resistance Mutation Identification. Mutations were filtered using the following parameters: have a negative effect on the majority of drugs in a single class; occurrence in the data set more than once; average Log FC for all compounds > 0.4; standard deviation over this average < 0.4. This provided us with a number of mutations that lead to an increase in fold change on average, again using literature we discarded any previously known mutations and kept those mutations that were novel.[6, 18, 40]

6.5.11 Drug Specific Resistance Mutation Identification. For all interpretable mutations, the standard deviation was calculated over the average Log FC values per drug within a class. Subsequently all mutations were ranked and the top 30 were retained here. The goal here was to find mutations that have the most diverse effect over the different drugs within a class.

6.5.12 Benchmark dataset for sequence only model comparison. The dataset we used to compare the performance of PCM models with sequence only models was obtained from Van der Borght *et al.*[39] From the paper the 150 sequences with the largest prediction error were selected per drug class. For mixtures present in this set the average value of each z-scale for each of the present variants at a single position was used as descriptor. Mixtures with more than four possible variants at a single position were discarded leading to a total of 146 NNRTI sequences, 146 NRTI sequences, and 149 PI sequences.

6.5.13 Stanford University Validation Set. Prediction of the Stanford University set is of is of particular interest since the correlation between Phenosense and AVG has previously been shown not to be very strong.[63, 64] In particular for mutations M41L, M184V, and T215Y there are differences in Phenosense predictions compared with AVG.[65] While the correlation between Phenosense and VircoTYPE (trained on AVG) is slightly better, there are discrepancies. For instance the resistance profile of d4T and TDF, have been shown to have a Pearson's correlation coefficient < 0.8 between the two assays.[51]

The reference set was downloaded from the Stanford website (version 5.0, July 30, 2010), from this set the sequences by Virco were removed (as they are presumed to be in the training set, and this would artificially boost the results). The mixtures were removed and this provided us with the following numbers of sequence – compound pairs: 1,252 (NNRTI), 2,190, (NRTI), and 4,356 (PI).

After we predicted the Log FC values for individual drug – mutant pairs using our models, the validation parameters were calculated grouped by: Sequence ID (average and standard deviation), per Isolate (average and standard deviation), per Reference ID (average and standard deviation), per drug (average and standard deviation), per class (total), and per Drug (total) (**Table 6.5**). The predictions per class are also included in **Figure 6.6**. Note that the raw data was used and no selection for high quality data was made, furthermore, the data was gathered at different labs, using different assays.

6.5.14 Clinical Cut-offs. Resistance was also classified using clinical cut-offs (CCOs), here we used the values provided on the Stanford website and the values from AVG were obtained from Virco (supporting **Table S5** –supporting **Table S7**). Subsequently CCP was calculated as a fraction of the total, in addition the fraction of overpredicted clinical response (resistance is predicted higher than measured experimentally) and underpredicted clinical response (resistance is predicted lower than measured experimentally) is included.

6.6 Supporting Information

Additional tables (supporting **Tables S1 – S11**), figures (**Figures S1 – S18**) are available as pdf. These materials are available online at www.gjpvanwesten.nl.

6.7 Acknowledgements

The authors would like to thank Koen van der Borght for providing the benchmark dataset for comparison with sequence only models.

6.8 References

1. F. Barre-Sinoussi, J. Chermann, et al.; *Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS).* Science; *1983*. **220** (4599): 868-871.
2. M. Popovic, M. Sarngadharan, et al.; *Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS.* Science; *1984*. **224** (4648): 497-500.
3. UNAIDS. *Progress Report: Global HIV/AIDS Response*. 2011 [cited 2012 January 30]; Available from: http://whqlibdoc.who.int/publications/2011/9789241502986_eng.pdf.
4. D. Kaufmann, G. Pantaleo, et al.; *CD4-cell count in HIV-1-infected individuals remaining viraemic with highly active antiretroviral therapy (HAART). Swiss HIV Cohort Study.* Lancet; *1998*. **351** (9104): 723-724.
5. F.J. Palella, K.M. Delaney, et al.; *Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection.* N. Engl. J. Med.; *1998*. **338** (13): 853-860.
6. V.A. Johnson, V. Calvez, et al.; *2011 update of the drug resistance mutations in HIV-1.* Topics in Antiviral Medicine; *2011*. **19** (4): 156-164.
7. B. Preston, B. Poiesz, and L. Loeb; *Fidelity of HIV-1 reverse transcriptase.* Science; *1988*. **242** (4882): 1168-1171.
8. J. Roberts, K. Bebenek, and T. Kunkel; *The accuracy of reverse transcriptase from HIV-1.* Science; *1988*. **242** (4882): 1171-1173.
9. K. Hertogs, S. Bloor, et al.; *Phenotypic and genotypic analysis of clinical HIV-1 isolates reveals extensive protease inhibitor cross-resistance: a survey of over 6000 samples.* AIDS; *2000*. **14** (9): 1203-1210.
10. C.C.J. Carpenter, D.A. Cooper, et al.; *Antiretroviral Therapy in Adults.* JAMA: The Journal of the American Medical Association; *2000*. **283** (3): 381-390.
11. L. Perrin and A. Telenti; *HIV Treatment Failure: Testing for HIV Resistance in Clinical Practice.* Science; *1998*. **280** (5371): 1871-1873.
12. K. Hertogs, M.-P. de Bethune, et al.; *A Rapid Method for Simultaneous Detection of Phenotypic Resistance to Inhibitors of Protease and Reverse Transcriptase in Recombinant Human Immunodeficiency Virus Type 1 Isolates from Patients Treated with Antiretroviral Drugs.* Antimicrob. Agents Chemother.; *1998*. **42** (2): 269-276.
13. C.J. Petropoulos, N.T. Parkin, et al.; *A Novel Phenotypic Drug Susceptibility Assay for Human Immunodeficiency Virus Type 1.* Antimicrob. Agents Chemother.; *2000*. **44** (4): 920-928.

14. H. Walter, B. Schmidt, et al.; *Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors.* Journal of Clinical Virology; *1999*. **13** (1–2): 71-80.

15. K. Van Laethem, A. De Luca, et al.; *A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients.* Antiviral therapy; *2002*. **7** (2): 123-129.

16. A. De Luca, A. Cingolani, et al.; *Variable Prediction of Antiretroviral Treatment Outcome by Different Systems for Interpreting Genotypic Human Immunodeficiency Virus Type 1 Drug Resistance.* J. Infect. Dis.; *2003*. **187** (12): 1934-1943.

17. J.-L. Meynard, M. Vray, et al.; *Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial.* AIDS; *2002*. **16** (5): 727-736.

18. Robert W. Shafer; *Rationale and Uses of a Public HIV Drug-Resistance Database.* J. Infect. Dis.; *2006*. **194** (Supplement 1): S51-S58.

19. A. Altmann, M. Däumer, et al.; *Predicting the Response to Combination Antiretroviral Therapy: Retrospective Validation of geno2pheno-THEO on a Large Clinical Database.* J. Infect. Dis.; *2009*. **199** (7): 999-1006.

20. N. Beerenwinkel, M. Daumer, et al.; *Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes.* Nucleic Acids Res.; *2003*. **31**: 3850 - 3855.

21. M.J. Perez-Elias, I. Garcia-Arota, et al.; *Phenotype or virtual phenotype for choosing antiretroviral therapy after failure: a prospective, randomized study.* Antiviral therapy; *2003*. **8** (6): 577-584.

22. L. Ratner, W. Haseltine, et al.; *Complete nucleotide sequence of the AIDS virus, HTLV-III.* Nature; *1985*. **313** (6000): 277-284.

23. B.T. Korber, B.T. Foley, et al. *Numbering Positions in HIV Relative to HXB2CG*. *1998*.

24. A. Adachi, H.E. Gendelman, et al.; *Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone.* J. Virol.; *1986*. **59** (2): 284-291.

25. H. Vermeiren, E. Van Craenenbroeck, et al.; *Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling.* J. Virol. Methods; *2007*. **145** (1): 47-55.

26. A. DiRienzo, V. DeGruttola, et al.; *Non-parametric methods to predict HIV drug susceptibility phenotype from genotype.* Stat Med; *2003*. **22**: 2785 - 2798.

27. H.C. Lim, M.E. Curlin, and J.E. Mittler; *HIV Therapy Simulator: a graphical user interface for comparing the effectiveness of novel therapy regimens.* Bioinformatics; *2011*. **27** (21): 3065-3066.

28. N. Beerenwinkel, T. Lengauer, et al.; *Methods for optimizing antiviral combination therapies.* Bioinformatics; *2003*. **19** (suppl 1): i16-i25.

29. A. Bender and R.C. Glen; *Molecular similarity: a key technique in molecular informatics.* Org. Biomol. Chem.; *2004*. **2**: 3204-3218.

30. M. Lapins, M. Eklund, et al.; *Proteochemometric modeling of HIV protease susceptibility.* BMC Bioinformatics; *2008*. **9** (1): 181-192.

31. A. Kontijevskis, R. Petrovska, et al.; *Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates.* Bioorg. Med. Chem.; *2009*. **17** (14): 5229-5237.

32. G.J.P. Van Westen, J.K. Wegner, et al.; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.

33. G.J.P. Van Westen, J.K. Wegner, et al.; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

34. M. Lapins and J.E.S. Wikberg; *Proteochemometric Modeling of Drug Resistance over the Mutational Space for Multiple HIV Protease Variants and Multiple Protease Inhibitors.* J. Chem. Inf. Model.; *2009*. **49** (5): 1202-1210.

35. M. Junaid, M. Lapins, et al.; *Proteochemometric Modeling of the Susceptibility of Mutated Variants of the HIV-1 Virus to Reverse Transcriptase Inhibitors.* PLoS One; *2010*. **5** (12): e14353.

36. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.

37. A. Tropsha and A. Golbraikh; *Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening.* Curr. Pharm. Des.; *2007*. **13** (34): 3494-3504.

38. D.E. Patterson, R.D. Cramer, et al.; *Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors.* J. Med. Chem.; *1996*. **39** (16): 3049-3059.

39. K. Van der Borght, E. Van Craenenbroeck, et al.; *Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations.* Bmc Bioinformatics; *2011*. **12** (1): 386.

40. V. Johnson, F. Brun Vezinet, et al.; *Update of the drug resistance mutations in HIV-1: Fall 2006.* Topics in HIV medicine; *2006*. **14** (3): 125-130.

41. J. Vingerhoets, M. Peeters, et al.; *An update of the list of NNRTI mutations associated with decreased virological response to etravirine: multivariate analysis on the pooled DUET-1 and DUET-2 clinical trial data [abstract 24].* Antiviral therapy; *2008*. **13**: Suppl 3:A26.

42. R.W. Shafer and J.M. Schapiro; *HIV-1 drug resistance mutations: an updated framework for the second decade of HAART.* AIDS reviews; *2008*. **10** (2): 67-84.

43. C.F. Perno, V. Svicher, and F. Ceccherini-Silberstein; *Novel drug resistance mutations in HIV: recognition and clinical relevance.* AIDS reviews; *2006*. **8** (4): 179-190.

44. S.-Y. Rhee, J. Taylor, et al.; *Genotypic predictors of human immunodeficiency virus type 1 drug resistance.* Proceedings of the National Academy of Sciences; *2006*. **103** (46): 17355-17360.

45. E.A. Emini, D.J. Graham, et al.; *HIV and multidrug resistance.* Nature; *1993*. **364** (6439): 679-679.

46. M. Tisdale, S.D. Kemp, et al.; *Rapid in vitro selection of human immunodeficiency virus type 1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse transcriptase.* Proceedings of the National Academy of Sciences; *1993*. **90** (12): 5653-5656.

47. V.W. Byrnes, E.A. Emini, et al.; *Susceptibilities of human immunodeficiency virus type 1 enzyme and viral variants expressing multiple resistance-engendering amino acid substitutions to reserve transcriptase inhibitors.* Antimicrob. Agents Chemother.; *1994*. **38** (6): 1404-1407.

48. T. Imamichi, T. Sinha, et al.; *High-Level Resistance to 3'-Azido-3'-Deoxythimidine due to a Deletion in the Reverse Transcriptase Gene of Human Immunodeficiency Virus Type 1.* J. Virol.; *2000*. **74** (2): 1023-1028.

49. S. Paolucci, F. Baldanti, et al.; *Gln145Met/Leu Changes in Human Immunodeficiency Virus Type 1 Reverse Transcriptase Confer Resistance to Nucleoside and Nonnucleoside Analogs and Impair Virus Replication.* Antimicrob. Agents Chemother.; *2004*. **48** (12): 4611-4617.

50. A.R. Zolopa, L.C. Lazzeroni, et al.; *Accuracy, Precision, and Consistency of Expert HIV Type 1 Genotype Interpretation: An International Comparison (The GUESS Study).* Clin. Infect. Dis.; *2005*. **41** (1): 92-99.

51. M. Van Houtte, G. Picchio, et al.; *A comparison of HIV-1 drug susceptibility as provided by conventional phenotyping and by a phenotype prediction tool based on viral genotype.* Journal of Medical Virology; *2009*. **81** (10): 1702-1709.

52. M. Sandberg, L. Eriksson, et al.; *New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids.* J. Med. Chem.; *1998*. **41** (14): 2481-2491.

53. S. Hellberg, M. Sjoestroem, et al.; *Peptide quantitative structure-activity relationships, a multivariate approach.* J. Med. Chem.; *1987*. **30** (7): 1126-1135.

54. R.C. Glen, A. Bender, et al.; *Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME.* IDrugs; *2006*. **9** (3): 199 - 204.

55. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.

56. Accelrys Software Inc *Pipeline Pilot Student Edition* Scitegic Version 6.1.5

57. A. Bender, J.L. Jenkins, et al.; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.

58. P. Willett, J.M. Barnard, and G.M. Downs; *Chemical Similarity Searching.* J. Chem. Inf. Comput. Sci.; *1998*. **38** (6): 983-996.

59. R Development Core Team; *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; *2006*.

60. E. Dimitriadou, K. Hornik, et al. *Misc Functions of the Department of Statistics (e1071)* TU Wien 2006 1.5-15

61. V. Vapnik; *The Nature of Statistical Learning1995*; New York: Springer.

62. A. Golbraikh and A. Tropsha; *Beware of q2!* Journal of Molecular Graphics and Modelling; *2002*. **20** (4): 269-276.

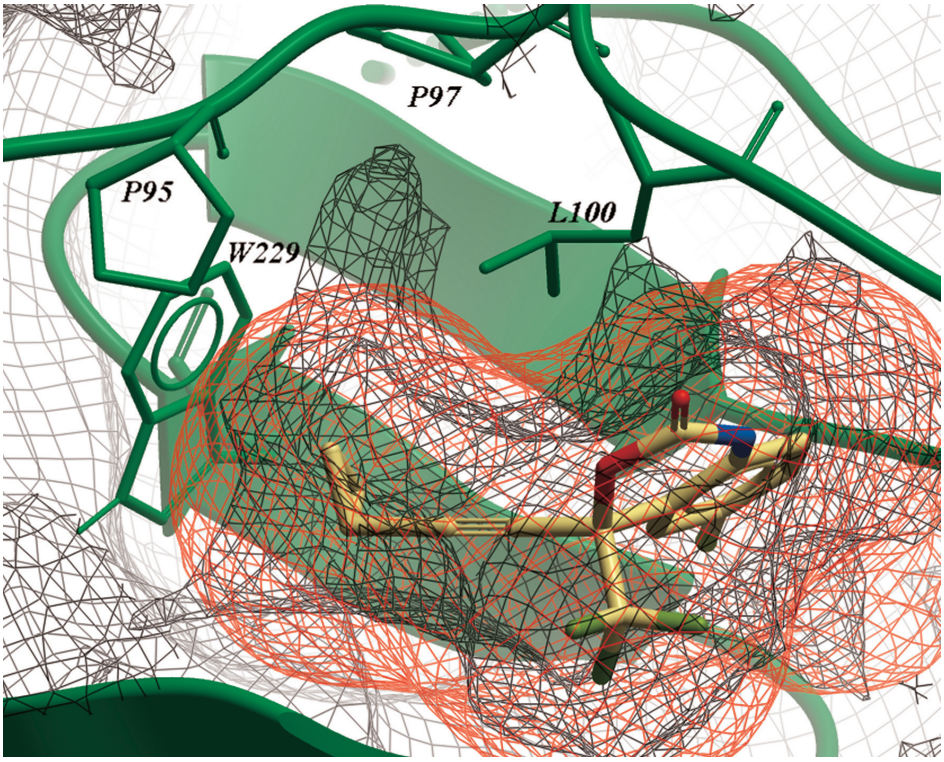
63. K. Wang, R. Samudrala, and J. Mittler; *Weak Agreement between Antivirogram and PhenoSense Assays in Predicting Reduced Susceptibility to Antiretroviral Drugs.* J. Clin. Microbiol.; *2004*. **42** (5): 2353-2354.

64. K. Wang, R. Samudrala, and J. Mittler; *Antivirogram or phenosense: a comparison of their reproducibility and an analysis of their correlation.* Antiviral therapy; *2004*. **9** (5): 703-712.

65. J. Zhang, S.-Y. Rhee, et al.; *Comparison of the Precision and Sensitivity of the Antivirogram and PhenoSense HIV Drug Susceptibility Assays.* JAIDS Journal of Acquired Immune Deficiency Syndromes; *2005*. **38** (4): 439-444.

Chapter 7

Mining Protein Dynamics from Sets of Crystal Structures using 'Consensus Structures'



G.J.P. van Westen, J.K. Wegner, A. Bender, A.P. IJzerman, and H.W.T. van Vlijmen; Protein Sci.; 2010. ***19*** *(4): 742-752.*

Contents

7.1 Abstract

In this work we describe two novel approaches to utilize the dynamic structure information implicitly contained in large crystal structure data sets. The first approach visualizes both consistent as well as variable ligand-induced changes in ligand-bound compared to apo protein crystal structures. For this purpose, information was mined from B-factors and ligand-induced residue displacements in multiple crystal structures, minimizing experimental error and noise. With this approach, the mechanism of action of non-nucleoside reverse transcriptase inhibitors (NNRTIs) as an inseparable combination of distortion of protein dynamics and conformational changes of HIV-1 reverse transcriptase was corroborated (a combination of the previously proposed 'molecular arthritis' and 'distorted site' mechanisms). The second approach presented here uses 'consensus structures' to map common binding features that are present in a set of structures of NNRTI-bound HIV-1 reverse transcriptase. Consensus structures are based on different levels of structural overlap of multiple crystal structures, and are used to analyze protein-ligand interactions. The structures are shown to yield information about conserved hydrogen bonding interactions as well as binding-pocket flexibility, shape and volume. From the consensus structures, a common wild type NNRTI binding pocket emerges. Furthermore, we were able to identify a conserved backbone hydrogen bond acceptor at P236 and a novel hydrophobic subpocket which are not yet utilized by current drugs. Our methods introduced here reinterpret the atom information and make use of the data variability by using multiple structures, complementing classical 3D structural information of single structures.

7.2 Introduction

The availability of crystal structures in both public archives (such as the Protein Data Bank (PDB) [1]) as well as proprietary repositories (such as within pharmaceutical companies) is growing at a phenomenal speed.[1] Crystal structures can provide a wealth of experimental data to the scientist, but the information obtained is static and cannot accurately depict the actual dynamic properties of the protein and its ligand.[2] Additional information that can provide insights into the dynamics is implicitly contained within a larger group of crystal structures of the same protein, as this set of structures captures (part of) the dynamically accessible conformation space of the protein. The challenge resides in how to mine this wealth of data. In the work presented here we will introduce two different methods for mining large sets of ligand-protein crystal structures.

7.2.1 Signal from noise. Our first approach mines data from B-factor values and ligand-induced residue displacements. In a single structure, information concerning the dynamics is provided by B-factors, which reflect the fluctuations of atoms around their average position in the crystal.[3] However, B-factors are also influenced by experimental error, temperature and crystal quality. Therefore it is per se difficult to distinguish this dynamic information from measurement errors and artifacts in situations where only a single structure is studied. The utilization of multiple structures can alleviate this problem,[4] provided that the B-factor values are normalized before comparing different structures.[5]

A closely related second approach is mapping ligand-induced changes in residue orientation by comparing apo structures with ligand bound structures.[6] Similarly, when only one pair of structures, i.e. one apo structure and one ligand-bound structure are compared, it is difficult to distinguish useful information from experimental artifacts. Our hypothesis governing the current work was that the simultaneous analysis of several apo and ligand-bound structures will lead to a better understanding of information common to all structures, highlighting trends and distinguishing them from artifacts or noise.

7.2.2 Consensus Structures. The third approach is to analyze the common spatial and pharmacophoric interaction properties of the available crystal structures, which we named 'Consensus Structures'. Existing approaches to derive a consensus structure have been aimed at mapping common features of a group of known ligands and creating a consensus pharmacophore.[7-9]

However, this ligand-based approach does not take any protein information into account, resulting in the inability of such approaches to extract protein-related information. Furthermore, it has already been shown that consensus information derived from several protein crystal structures can indeed extrapolate beyond the original data.[10, 11] Consensus structures are based on the aligned ligand binding pockets of multiple ligand-bound crystal structures and allow analysis of the shape and pharmacophoric patterns present in all of the structures, as well as the differences between them. Consensus structures combine information about different binding site geometries to identify key features responsible for ligand binding. Isocontour consensus surfaces that visualize features common to a minimum percentage of the total structures used are obtained from these Consensus structures and allow visualization of the degree of conservation of the protein or protein features.

7.2.3 Case Study. In a case study, we applied the above methods to a data set consisting of human immunodeficiency virus type-1 reverse transcriptase (HIV-1 RT) structures in complex with non-nucleoside reverse transcriptase inhibitors (NNRTIs). HIV-1 RT is one of the most studied drug targets known today and was the first target identified in the treatment of infection with HIV-1.[12] As a result, a large number of crystal structures are available in the PDB, rendering this target suitable for our first case study.[1] NNRTIs are non-competitive inhibitors of HIV-1 RT acting on an allosteric binding pocket with high specificity. However, the nature of HIV-1 replication leads to a quick onset of resistance of HIV-1 towards NNRTIs.[13, 14] This resistance forms an increasing problem in the treatment of HIV-1 infection and is mainly caused by point mutations in the protein.[13-16] HIV-1 RT is a heterodimer, consisting of a large 560-residue subunit (p66) and a smaller 440-residue subunit (p51). The catalytic site on the p66 unit consists of a conserved YMDD motif and a third aspartic acid (residues D110, Y183, M184, D185 and D186). The rather flexible pocket is not present in the apo form of the enzyme and is only created upon binding of an NNRTI to HIV-1 RT, thus reducing enzyme flexibility.[17] Furthermore, it has been shown that the flexibility of HIV-1 RT depends on its ligation state, and is increased upon DNA binding.[18] The information mined from the B-factors and ligand-induced changes of the HIV-RT crystal structures enabled us to explore the mechanism of NNRTI inhibition in more detail. The consensus structure analysis resulted in the identification of conserved hydrophobic and hydrogen bonding features that provided new insights and design options for HIV-1 RT inhibition by NNRTIs.

7.3 Results and discussion

7.3.1 B-factor analysis. Firstly, normalized B-factors of NNRTI-bound enzymes were compared with the corresponding values in apo enzymes. Our results show a significant decrease in B-factors of the entire pocket upon NNRTI binding, indicating smaller fluctuations and a stiffer protein backbone in this region. This reduction in flexibility is in agreement with earlier MD simulations.[17, 19] While B-factors of most residues respond variably upon ligand binding (see **Figure 7.1**), the loop containing two of the catalytic site residues (D185 and D186) and the neighboring residues, i.e. residues 181 to 188, shows a significant decrease in flexibility (see also supporting **Figure S1**).

The restriction of conformational change of this loop by NNRTIs, which was proposed to be the mechanism of action by Das *et al.*,[20] is fully supported by our results on the significantly decreased B-factors of this region. In contrast, the region between residues L228 and L234 undergoes a consistent average increase in flexibility. This region contains the 'primer grip' residue M230 and the increase would be unfavorable for its function retaining the growing DNA strand. A principal component analysis (PCA) of the normalized B-factor distributions over the residues shows that all structures form a cluster which shows significant diversity (but no outliers) and that the three apo structures are nearest neighbors (supporting **Figure S2**). A heat map of the normalized B-factor profiles of all structures can be found in supporting **Figure S3**. To compare our findings to DNA-bound protein, all crystal structures were compared to three different DNA-bound structures.

However, contrary to the consistent profile observed in the ligand bound or apo structures the overall B-factor profile differs between these DNA-bound structures. We therefore do not have enough data to draw firm conclusions on flexibility changes upon DNA binding. When we separated all NNRTI-bound structures into wild type (wt) and mutated structures a trend was observed that resistance-conferring point mutations lead to a partial restoration of flexibility of the catalytic site region. (Supporting **Figure S1**) This trend is consistent with the findings of Zhou *et al.*[17]

This type of flexibility information is useful when implementing protein flexibility in e.g. docking calculations.[21, 22] Awareness about the flexible residues enables focusing computational expense on only these parts of the protein, while still allowing induced fit to take place to the required degree.

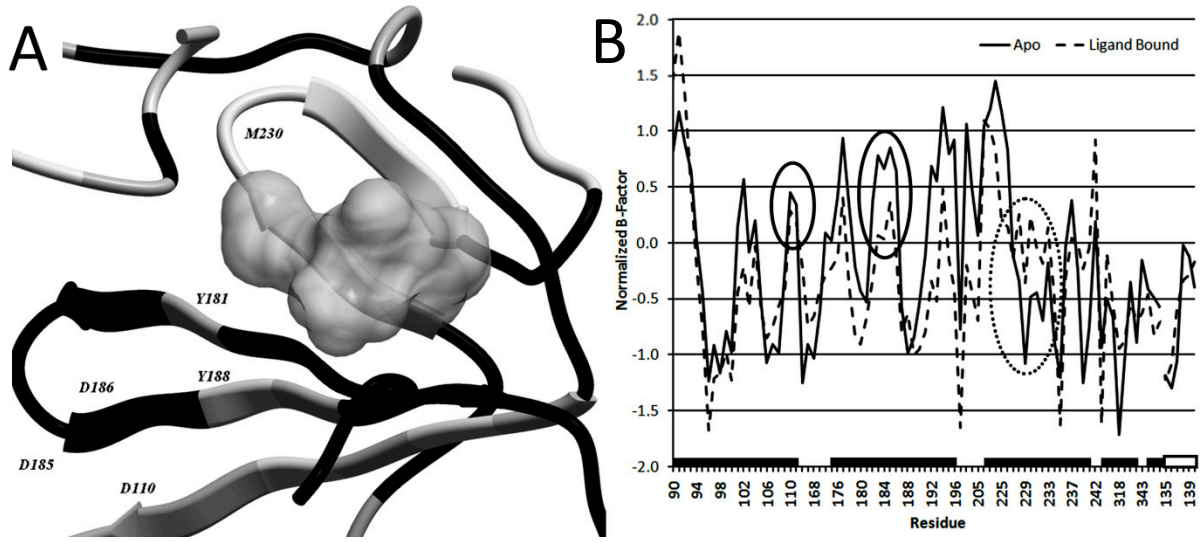


Figure 7.1: The backbone of the NNRTI pocket, colored by the changes in average B-factor. White residues indicate an increase (≥ 0.2), grey indicates no change (between 0.2 and -0.2) and black indicates a decrease (≤ -0.2). (A, PDB 1FK9) The grey volume represents the NNRTI Efavirenz. The entire pocket region has a lower average normalized B-factor in the ligand-bound compared to the apo form while the profile remains comparable (B). The catalytic residues undergo a decrease in B-factor (black circles), while the primer grip region containing residue W229 and M230 undergoes an increase upon ligand binding (dashed black circle). The black bars on the horizontal axis indicate continuous residues, the white filled bar indicates the residues on the p51 subunit. Each tick marks a separate residue the precise residue numbers can be found in the materials and methods, p66 and p51 have been separated by a gap.

7.3.2 Ligand induced displacement analysis. We next analyzed the residue displacements resulting from ligand binding. This analysis confirms that all NNRTIs induce a similar binding pocket into the protein when compared to the apo structure. This holds true for both the backbone carbon alpha atoms as well as the movement of the center of mass of the residues (supporting **Figures S4-S11**). The only exception is observed in the Capravirine structure; however, as we only had one structure available we cannot confirm this to be an effect characteristic for the ligand or for this particular crystal structure.

Hence, this structure was removed from further analysis. Next, the residue center of mass absolute displacement distances were examined. **Figure 7.2A** depicts the pocket backbone colored according to the values obtained from the calculation of the displacement upon ligand binding. A PCA analysis of the displacement vectors shows that the ligand-bound structures cluster together, even more than do the apo structures (supporting **Figure S4**), confirming the formation of a common pocket upon ligand binding.

Summarizing our findings for NNRTI binding, firstly the shift of the catalytic acid residues in **Figure 7.2B** is conserved. Secondly, the known shifts as a result of the flip of residues Y181, Y188 and the known upward movement of W229 and M230 are confirmed for all NNRTIs.[23] The entire ß12 sheet (P225 – P236) known to be in contact with the growing DNA template via residue M230,[20, 24-26] is displaced upwards into the DNA binding groove, adapting to the ligand upon binding. However, the displacement of the sheet is a rotation pivoting around residues P226 and L234/H235, which remain relatively in place. This suggests how the sheet adapts itself to the size of the bound NNRTI while moving the primer grip away from the nucleotide binding region. The mutations present in the mutated RT structures do not appear to have a significant influence on residue movement. (For detailed heat maps indicating the shift per residue see supporting **Figures S6-S10**.) Interestingly, the distance measured does not necessarily correlate with these residues being involved in resistance mutations. However, it is striking that some of the residues that are detected to move relatively large distances in all structures are in known locations for resistance conferring mutations. Because these residues undergo the largest changes resulting from NNRTI binding, changes occurring in these residues are most likely to perturb NNRTI binding.

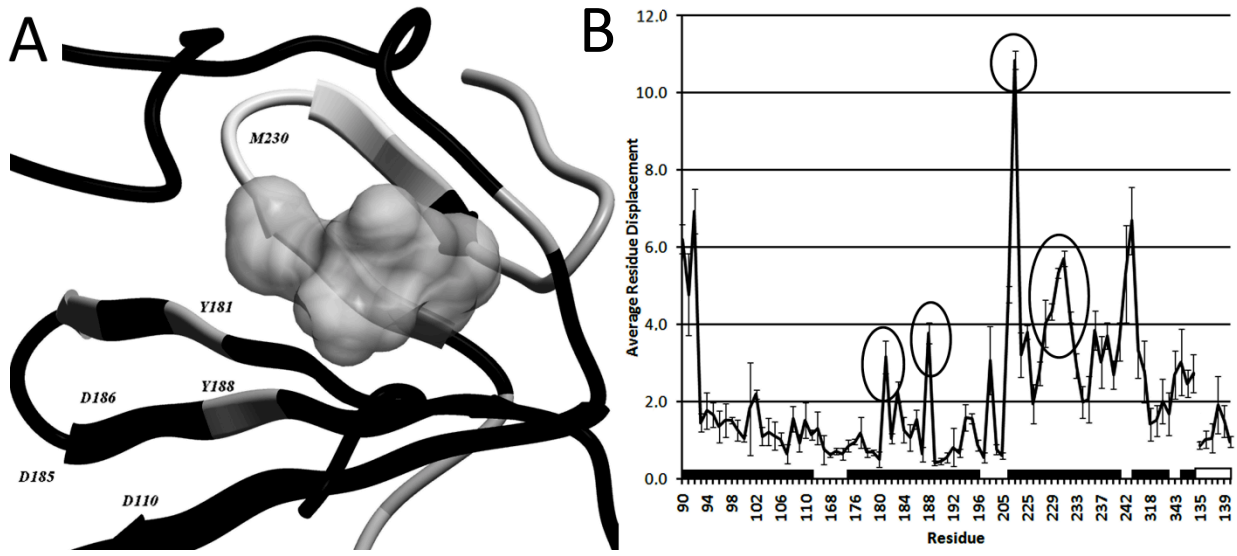


Figure 7.2: The backbone of the NNRTI pocket, colored by the average residue displacement. Black indicates a small average displacement, ≤ 2 Å, grey indicates a medium displacement, between 2 and 4 Å, and white indicates a large displacement ≥ 4 Å (A). The grey volume represents the NNRTI Efavirenz. Residues Y181, Y188, K223 and M230 all undergo a relatively large movement upon ligand binding (B, black circles). The residues between L228 and L234 all undergo a large displacement upon ligand binding, this correlates with the results from the B-factor analysis. The residue axis is labeled identically to **Figure 7.1**.

Upon DNA binding, the position of the catalytic site loop carrying residues Y181 through Y188 is consistently shifted (**Figure 7.3**). The loop maintains the overall conformation suggesting a hinge-like movement during catalysis. Furthermore, the known shift of the catalytic loop induced by NNRTI binding is in fact opposite to the path of the downward movement induced by DNA binding. (Supporting **Figures S6 – S11**) We compared the shift on each cartesian axis as a result of either DNA binding or NNRTI binding. The figures clearly show that the catalytic loop and especially residues 184-186 move in an opposite direction when compared to the apo structures in both the case of Cα and centroid movement on both the x- and z-axes. In the case of the y-axis the result is less pronounced.

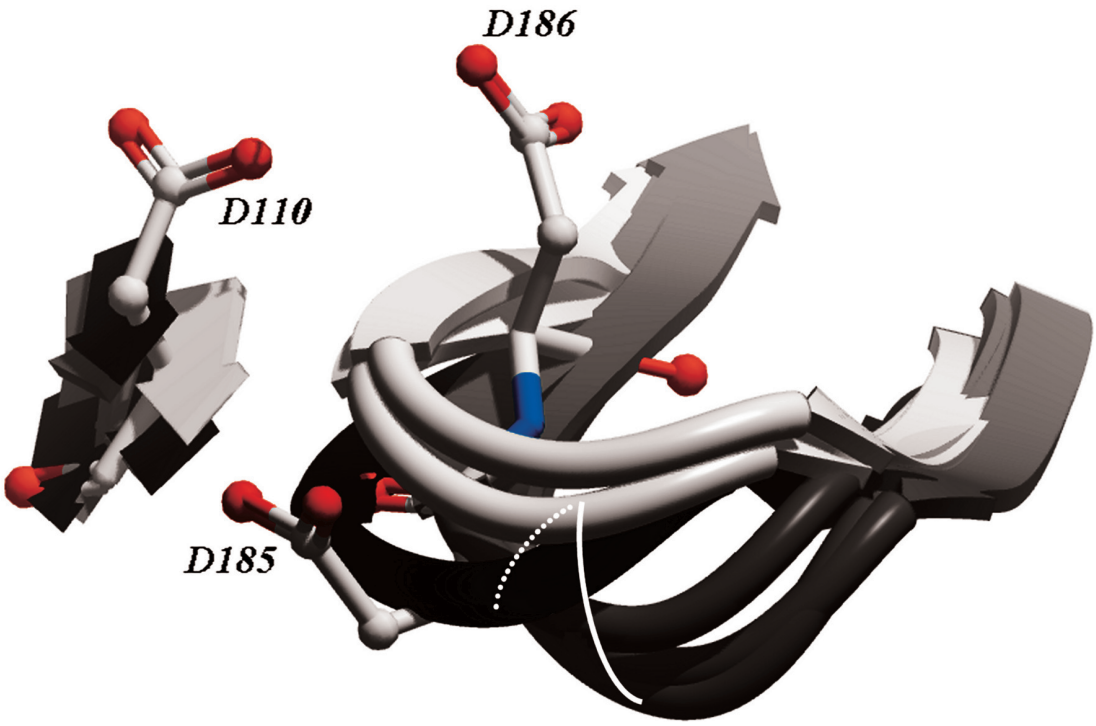


Figure 7.3: Overview of the changes occurring at the catalytic site as a result of DNA binding (three grey ribbons), NNRTI binding (two black ribbons) compared to the apo position (two white ribbons). The three aspartic acids that are visible are part of the DNA bound conformation. (1RTD) In the apo structures, these residue side chains are placed similarly, however in the NNRTI bound structures D186 points toward D110, D110 points outward and D185 points downward. Upon DNA binding the loop containing the residues moves along the path indicated by the white curve, in the presence of an NNRTI the loop moves along the path indicated by the dashed white curve.

7.3.3 NNRTI working mechanism. Combining the above findings on the movement and flexibility of HIV-RT residues upon binding of an NNRTI allows us to elaborate on previously proposed working mechanisms of NNRTIs. The working mechanism has been proposed to be either the result of a catalytic site distortion,[26-30] or a more rigid protein.[19, 29, 30] From our flexibility analysis we observe that the binding of an NNRTI to HIV-1 RT leads to a shift in flexibility of residues around the binding site, with mobile residues become more rigid while rigid residues become more mobile. In addition, upon NNRTI binding the catalytic triad residues and their neighboring residues, D110-V111 and Y181-Y188, but also the primer grip region, M230 and surrounding residues, undergo a consistent displacement opposite to the displacement induced by DNA binding. As a result the conformation of the catalytic loop is distorted, enlarging the distance between primer grip and the nucleotide binding site. These changes were found to be consistent among all crystal structures studied. At the catalytic site, the 4 Å average movement of the aspartic acids is a large distance as the three-dimensional orientation of the catalytic motif and the nucleotide is crucial in catalyzing the reaction. Mendieta *et al.* have found that the Mg2+ ions stabilize the catalytic complex and lower the catalytic attack distance to a stable 3 Å.[31] These essential Mg2+ ions were missing in all crystal structures containing and NNRTI, indicating that the changed orientation of the aspartic acids might not be able to contain these ions.

The large movement of the primer grip away from the catalytic site upon NNRTI binding and the increase in flexibility are consistent among all structures. Both of these changes are likely to lead to a decreased reaction rate as they inhibit the function of retaining the DNA strand. We therefore conclude that the NNRTIs disrupt both protein conformation and dynamics and that it is this combination that inhibits the function of HIV-RT. Thus we propose a working mechanism for NNRTIs that is a combination of both the rigid protein, distorted-catalytic-site and distorted primer grip region theories. The bound NNRTI stabilizes the region surrounding the catalytic site in a conformation not able to perform catalysis. The effects on protein conformation and dynamics cannot be seen independently as one directly influences the other. Therefore it could be speculated that an increase in flexibility as a result of point mutations allows the primer grip and catalytic site to move closer together restoring catalytic activity. As a result this could lead to resistance of the particular HIV-RT mutant form. Our results support this theory.

When this manuscript was revised, Paris *et al.*[32] published a large scale comparison of NNRTI crystal structures. While their superposition is based on only a subset of residues and we left out the structures where an NNRTI was soaked out of the pocket, their main conclusions are generally in line with ours. Moreover, our results indicate that NNRTI binding influences not only primer grip movement, but also distortion of the catalytic site and changes in protein dynamics and that these are all in fact required for inhibitory activity.

7.3.4 Consensus structures. In the second part of our work on mining information from multiple crystal structures, we present the results from the consensus structures creation. The difference between two extreme situations of conservation values is visualized in **Figure 7.4**. The surface visualizing low conservation is shown as a green wire grid. This surface shows all parts of the three-dimensional space covered by at least 10 % of the protein structures, including different side-chain orientations. This surface features a rather large protein volume, and consequently the empty pocket volume in this map is smaller than in the high-conservation map. The low-conservation surface thus describes the unity of conformational space accessible to the protein in any one of the dataset structures. The surface visualizing high conservation is shown as a solid green mesh. This surface contains the volume covered by at least 50 % of the structures. It suggests the volume to which the protein binding pocket can be extended, describing the most conserved side-chain-occupied space.

This volume does not occur in any of the individual structures, and this high-conservation surface represents the largest possible binding pocket a hypothetical ligand could occupy. In addition, combining high and low conservation surfaces in one view can be used to locate regions of flexibility by visually comparing highly-conserved side-chain orientations, where there is little difference between high- and low-conservation surfaces, to side-chains that can move more freely, showing a large difference between high- and low-conservation surfaces.

When the consensus creation procedure is repeated on our set of NNRTIs, the surfaces visualize the space taken up by the different NNRTIs within HIV-1 RT. Here the low-conservation surface visualizes the unity of conformational space accessible to the NNRTIs in any one of the dataset structures, and a larger volume than present in any of the individual structures. The high-conservation surface visualizes the common volume that is used by all NNRTIs.

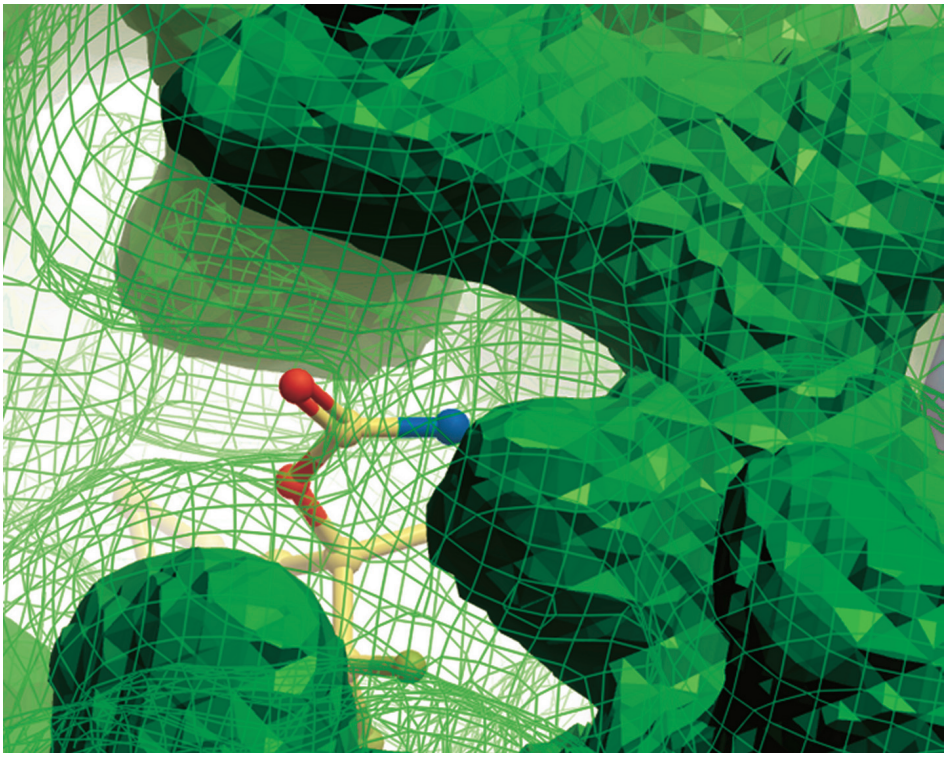


Figure 7.4: Difference between surfaces that represent low conservation and high conservation. All the grid points that are occupied in at least 10 % of the structures are shown as a green wire grid (low conservation points). All the grid points that represent at least 50 % of the structures are shown as a green solid mesh (high conservation points). Both represent an extreme value and are overlaid on the PDB structure 1FK9, depicted with its ligand Efavirenz. The wire grid surface shows all possible side chain locations (low conservation), while the solid mesh surface only shows the most conserved side chain location.

7.3.5 Consensus pocket. Van der Waals (VdW) consensus structures can visualize common features present in all structures, including crystal structures and optionally even homology models carrying point mutations.[33] The resulting consensus pocket shape can be regarded as a target shape for high affinity ligands since it represents the maximal theoretically accessible volume. As the highly flexible pocket adapts to each different NNRTI and there is no standardized wild-type pocket, consensus structures can provide this standardized pocket. Based on the degree of conservation of the surfaces, an estimate can be made of the space available for NNRTI binding and this can be related to the actual volume of NNRTIs. From **Table 7.1** it can be concluded that second-generation NNRTIs, Rilpivirine, Capravirine and Delavirdine, make better use of the maximal theoretical available space (66%, 69% and 75% respectively). Given our consensus structures, even larger NNRTIs than Delavirdine are theoretically possible. Coincidentally, while finalizing this manuscript, this has been experimentally confirmed by Sweeney *et al.* [34]

Table 7.1: Volumetric information relating the size of the consensus structures to the size of NNRTIs.

Volumetric object:	Volume (Å^3)
NNRTI consensus (50 % occupancy)	103
NNRTI consensus (30 % occupancy)	296
NNRTI consensus (10 % occupancy)	578
739W94	267
Nevirapine	274
1051U91	275
Efavirenz	305
HEPT	311
Alpha-Apa	315
PETT2	315
PETT1	333
HBY097	334
MKC442	336
UC781	341
9-chloro-TIBO	344
8-chloro-TIBO	344
TNK6123	364
Rilpivirine	379
TNK651	391
Capravirine	400
Delavirdine	432

Overview of the volumes contained within the different NNRTI consensus structures compared with the volume of NNRTIs in their bound confirmation as it was obtained from the crystal structure.

7.3.6 New pocket features. The combination of consensus structures of the pocket with consensus structures of the combined NNRTIs can pinpoint locations within the pocket that are not or inefficiently used by current drugs, shown in **Figure 7.5.** The figure was created combining a VdW surface of the binding pocket and a VdW NNRTI surface, both visualizing low conservation. A small conserved sub-pocket is visible in the low conservation protein surface, which means that it is present in all crystal structures. Located between residues P95, P97, L100 and W229 on the 'A' chain it might be the appropriate place to add an additional methyl group to a ligand to fill this lipophilic pocket. None of the NNRTIs from the studied crystal structures contacts the residues surrounding the sub-pocket, as illustrated for Efavirenz.

Analogously, VdW consensus structures can also be applied to other drug targets in the rational design of new ligands, leading to derivatives that occupy non-used pocket space, thus maximizing the contact surface and interaction potential.

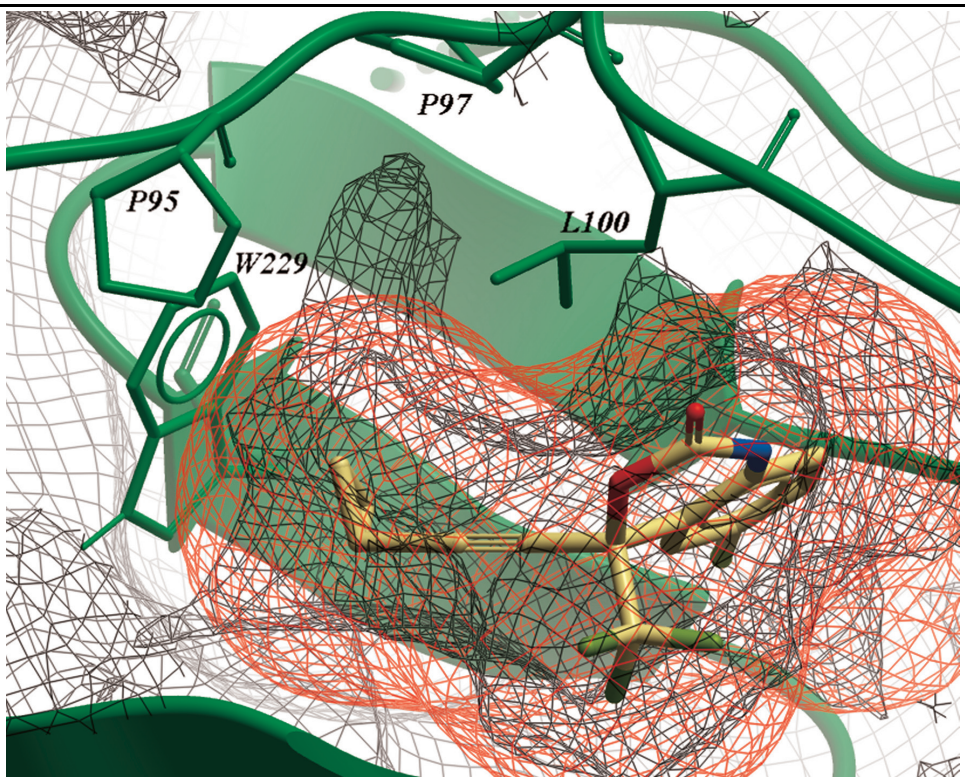


Figure 7.5: The consensus binding pocket. The protein VdW consensus surface is shown at 10 % occupancy (green wire grid) and is superimposed on the 10 % occupancy NNRTI VdW consensus surface (red wire grid). The PDB structure 1FK9 is shown in green and with yellow carbon atoms its ligand Efavirenz. Located between the residues P95, P97, L100, and W229 is a sub-pocket unused by current NNRTIs. This sub-pocket is in fact an extension of the known cavity located between Y181, Y188 and W229.

7.3.7 Hydrogen bonding information. The surfaces created based on hydrogen bonding (HB) occupancy visualize the conserved HB potential within the pocket or located on the NNRTIs, giving information about HB potential that can be exploited in drug discovery projects. Similar to VdW surfaces, variation of the conservation level identifies conserved HB locations. On our HIV-RT dataset for example, known backbone-NH HB interactions at K101 and K103 (**Figure 7.6A**) combined with the known backbone carbonyl HB at K101 and electrostatic interactions at H235 (**Figure 7.6B**) were confirmed using the consensus structures. Furthermore, a conserved carbonyl HB acceptor not used by NNRTIs can be identified at K102, although this acceptor is accessible in only a few of the ligand bound structures. Most interestingly we found that large NNRTIs that lead to a large shift of the ß12-sheet (P225 – P236) break the HB between backbone-NH at K103 and backbone carbonyl at P236. While the backbone-NH at K103 has been described to participate in HB interactions to NNRTIs,[35] the concurrently available carbonyl at P236 has not. Use of this HB acceptor could provide an additional backbone interaction to an NNRTI, which might lead to an improved resistance profile.

As some viral strains carry a K101P mutation, which does not allow the backbone-NH HB to be formed,[36] it is of high importance to identify additional common interaction sites in the NNRTI pocket that can be exploited. Although a P236L mutant has been identified, this will not change the potential for a backbone HB interaction. Additionally, this mutant has been described as having low replication fitness.[37]

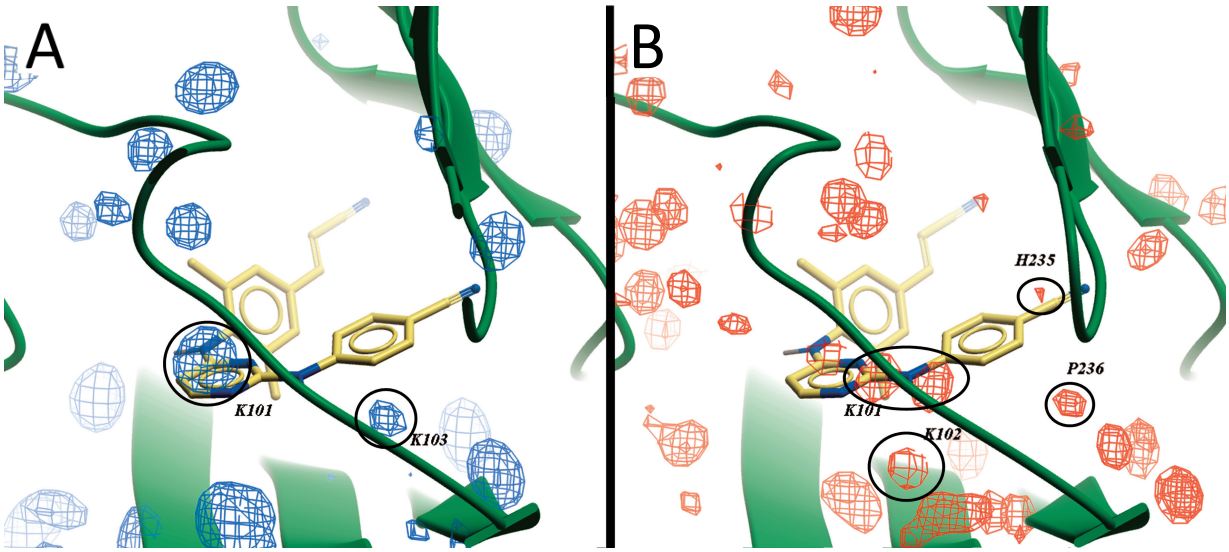


Figure 7.6: Consensus surfaces visualizing all HB locations. Both conserved HB donors surrounding the pocket (A) and all conserved HB acceptors (B) among the selection of crystal structures are shown. For reasons of clarity, only surfaces visualizing 20 % occupancy are shown. The three dimensional surfaces are superimposed on the PDB structure 2ZD1, the green ribbon indicating the backbone, and Rilpivirine depicted by yellow carbons.

7.4 Conclusion

In conclusion, both methods presented in this work, the analysis of B-factors and ligand induced residue displacement as well as the analysis of steric and pharmacophoric properties from multiple crystal structures, show how the growing number of crystal structures available can be mined efficiently to generate novel hypotheses for lead optimization. Although the amount of information available in these analyses increases with the number of structures included, the use of a handful of structures can already provide insights. Our findings were also supported by novel, more sterically-demanding NNRTIs that were published by Sweeney *et al.* just as this manuscript had been finalized.

Our methods facilitated the identification of the working mechanism of NNRTIs as a combination of two mechanisms that were previously suggested (the 'molecular arthritis' and 'distorted catalytic site' hypotheses). In addition, our consensus structures were able to extract conserved locations of interest from the crystal structures without the need for molecular dynamics. The different types of consensus structures can complement each other and provide a useful overview of the interaction between a class of compounds and its target protein. Using this method we identified a novel backbone hydrogen bond acceptor at P236 and a novel hydrophobic subpocket.

7.5 Materials and methods

7.5.1 Dataset. A total of 47 crystal structures were used in our analysis (**Table 7.2**). Crystal structures were obtained from the PDB and grouped according to bound NNRTI. The selection included several structures of the NNRTIs approved for clinical use, namely Nevirapine, Efavirenz, and Delavirdine. Several structures of apo HIV-1 RT have been used (group a) three of which contain a bound DNA fragment (group b). Apo structures 1JLE and 1RTJ were obtained by soaking out the NNRTI and are therefore not native apo structures. Accordingly, in an analysis of the structures, the orientation of the backbone was found to show more similarities with ligand-bound forms of HIV-RT than with the apo form. For this reason structures 1JLE and 1RTJ were omitted from our dataset. Out of the selected structures several contained a point mutation in the NNRTI pocket. From groups a and c in **Table 7.2**, B-factors were extracted in Molsoft ICM.[38] The consensus structures were created from a sub-selection of all 40 NNRTI-bound crystal structures, including 23 different NNRTIs (group c, **Table 7.2**). For all NNRTIs, the charge at pH 7, was calculated using the Marvin Beans pKa prediction tool by ChemAxon,[39] and all were found to be uncharged.

7.5.2 Computational details. All structural experiments and visualizations were performed using ICM. In all structures the NNRTIs were checked for inconsistencies and the bond orders were confirmed (A workflow of the performed experiments is given in supporting **Figure S12**). A multiple sequence alignment was created within ICM using default options, verifying that no mutations were present around the NNRTI pocket other than the known single mutations within the PDB structures. Hydrogens were added to the PDB by ICM object conversion, which contains hydrogen bond optimization, protonation state optimization of His residues, and rotamer optimization of Asn, Gln and His residues. Subsequently all RT crystal structures were superimposed based on alignment of the backbone atoms of selected residues. This selection was made using a 12 Å sphere around the largest NNRTI, Dlv, in PDB structure 1KLM. This included the following residues on chain A: 90-111, 161, 164, 168, 171, 172, 175-196, 198, 199, 201, 205, 222-240, 242, 315-321, 343, 348-350 and on chain B: 135–140. Visualizations and calculations were performed using the superimposed structures.

7.5.3 B-factor analysis. The influence of experimental error and conditions was minimized by normalizing the B-factors of every residue in the crystal structures similar to Yuan *et al*,[5] where we used full residue B-Factors instead of Cα-only B-Factors. This normalization uses the standard deviation over all B-factors per structure to allow a comparison of the values of different crystal structures.[5] After normalization, the B-factors from all ligand bound structures were combined and their mean value per residue was calculated, the same was done for the apo structures. Because of the very high temperature factors, structures 1HMV and 2ZE2 were omitted from these experiments. The average difference values between apo and ligand bound normalized B-factors per residue were binned in three classes: lower (≤ -0.2), virtually unchanged (between -0.2 and 0.2) and higher (≥ 0.2).

Table 7.2: Summary of the PDB structures that were used in all analyses.

PDBCode	Group	Mutation	Drug	Class	Resolution(Å)
1DLO	a	A172K	n/l	Apo	2.70
1HMV	a	n/p	n/l	Apo	3.20
1HQE	a	K103N	n/l	Apo	2.70
1QE1	a	M184I	n/l	Apo	2.85
1N6Q	b	n/p	n/l	DNA Bound	3.00
1RTD	b	n/p	n/l	DNA Bound	3.20
2HMI	b	n/p	n/l	DNA Bound	2.80
1RTH	c	n/p	1051U91	1051U91	2.20
1JLQ	c	n/p	739W94	739W94	3.00
1VRU	c	n/p	Alpha-Apa	Alpha-Apa	2.40
1EP4	c	n/p	Capravirine	Capravirine	2.50
2ZD1	c	n/p	Rilpivirine	DAPY	1.80
2ZE2	c	L100I/K103N	Rilpivirine	DAPY	2.90
3BGR	c	K103N/Y181C	Rilpivirine	DAPY	2.10
1KLM	c	n/p	Delavirdine	Delavirdine	2.65
1FK9	c	n/p	Efavirenz	Efavirenz	2.50
1FKO	c	K103N	Efavirenz	Efavirenz	2.90
1IKW	c	n/p	Efavirenz	Efavirenz	3.00
1JKH	c	Y181C	Efavirenz	Efavirenz	2.50
1BQM	c	n/p	HBY097	HBY097	3.10
1C1C	c	n/p	TNK6123	HEPT	2.50
1JLA	c	Y181C	TNK651	HEPT	2.50
1RT1	c	n/p	MKC442	HEPT	2.55
1RT2	c	n/p	TNK651	HEPT	2.55
1RTI	c	n/p	HEPT	HEPT	3.00
1FKP	c	K103N	Nevirapine	Nevirapine	2.90
1JLB	c	Y181C	Nevirapine	Nevirapine	3.00
1JLF	c	Y188C	Nevirapine	Nevirapine	2.60
1S1U	c	L100I	Nevirapine	Nevirapine	3.00
1S1X	c	V108I	Nevirapine	Nevirapine	2.80
1VRT	c	n/p	Nevirapine	Nevirapine	2.20
2HND	c	K101E	Nevirapine	Nevirapine	2.50
2HNY	c	E138K	Nevirapine	Nevirapine	2.50
1DTQ	c	n/p	PETT1	PETT	2.80
1DTT	c	n/p	PETT2	PETT	3.00
1JLC	c	Y181C	PETT2	PETT	3.00
1HNV	c	n/p	8TIBO	TIBO	3.00
1REV	c	n/p	9TIBO	TIBO	2.60
1TVR	c	n/p	9TIBO	TIBO	3.00
1UWB	c	Y181C	8TIBO	TIBO	3.20
1JLG	c	Y188C	UC781	UC	2.60
1RT4	c	n/p	UC781	UC	2.90
1RT5	c	n/p	UC10	UC	2.90

1RT6	c	n/p	UC38	UC	2.80
1RT7	c	n/p	UC84	UC	3.00
1S1T	c	L100I	UC781	UC	2.40
1S1W	c	V106A	UC781	UC	2.70

The table shows a summary of the PDB structures that were used in the analyses. All structures were subdivided into three groups; a, which contained apo-enzymes, b, which contained apo enzymes bound to a DNA fragment and c, which contained NNRTI bound structures. Furthermore mutations present in the NNRTI binding pocket were identified, if none were present 'n/p' was used. Structures marked with 'n/l' did not contain an NNRTI. Finally, all NNRTIs were subdivided into 13 classes as shown in the table.

7.5.4 Ligand induced displacement analysis. Ligand-induced conformational changes were characterized in three ways. Firstly, for each of the superimposed structures a scalar representing the mean distance for each pocket residue between the centroid in the ligand-bound conformation and the centroid of the two apo structures was determined. The obtained average centroid displacement distances were binned into three classes: small (≤ 2 Å), medium (between 2 Å and 4 Å) and large (≥ 4 Å). Secondly, the three-dimensional displacement vector between residue Cα positions in the ligand-bound conformation and the apo structures was split up over all three Cartesian axes to determine the backbone movement on each specific axis. Thirdly, the procedure was repeated between residue centroid positions in the ligand-bound conformation and the apo structures. Thus, the movement of the backbone and the side chains were both taken into account as well as changes in orientation of side chains (which are known to be relatively independent from backbone movements).[40]

7.5.5 Conversion of crystal structures to three-dimensional occupancy maps. The crystal structures were converted to three-dimensional occupancy maps by placing the structures in a grid box. A VdW occupancy value was assigned to each grid point and was normalized to a value between 0 and 1 using the internal auto trim function of ICM. This quasi-binary scaling enabled a comparison between occupancy maps obtained from different crystal structures.[41] For HB maps, the values were scaled between 0 and 1 for acceptors and between 0 and -1 for donors, similar to the VdW scaling.

7.5.6 Creation of consensus potentials and structures. Consensus structures were created from the initial superpositions by adding up the individual occupancy maps of the set of crystal structures. This step was performed separately for the RT binding pocket residues and the bound NNRTI structures. When several structures containing the same NNRTI were present, the average value of all structures containing that specific NNRTI was used. Thereby each NNRTI class contributed equally to the final occupancy map and the domination of a single class of NNRTIs over the others was avoided. Occupancy maps for the set of crystal structures were created for VdW, HB donors and HB acceptors, from the occupancy maps of each individual structure and named Consensus Structures. Isocontour surfaces were created at different levels, corresponding to visualization of different degrees of conservation.

7.6 Supporting Information

Additional Figures (**Figures S1 – S12**) are available online. These materials are available online at www.gjpvanwesten.nl.

7.7 Acknowledgements

We thank Andrew Orry (Molsoft L.L.C), Anik Peeters, Luc Geeraert, Ann Vos and Carlo Boutton (Tibotec-Virco BVBA) for their helpful discussions.

7.8 References

1. H.M. Berman, J. Westbrook, et al.; *The Protein Data Bank* Nucleic Acids Res.; *2000*. **28**: 235-242.
2. R.M. Knegtel, I.D. Kuntz, and C.M. Oshiro; *Molecular docking to ensembles of protein structures.* Journal of Molecular Biology; *1997*. **266** (2): 424-440.
3. Z. Yuan, T.L. Bailey, and R.D. Teasdale; *Prediction of protein B-factor profiles.* Proteins: Struct., Funct., Bioinf.; *2005*. **58** (4): 905-912.
4. J.E. Wampler; *Distribution Analysis of the Variation of B-Factors of X-ray Crystal Structures: Temperature and Structural Variations in Lysozyme.* J. Chem. Inf. Comput. Sci.; *1997*. **37** (6): 1171-1180.
5. Z. Yuan, J. Zhao, and Z.-X. Wang; *Flexibility analysis of enzyme active sites by crystallographic temperature factors.* Protein Eng.; *2003*. **16** (2): 109-114.

6. P.A. Keller, S.P. Leach, et al.; *Development of computational and graphical tools for analysis of movement and flexibility in large molecules.* Journal of Molecular Graphics and Modelling; *2000*. **18** (3): 235-241.

7. N. Richmond, C. Abrams, et al.; *GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D.* J. Comput.-Aided Mol. Des.; *2006*. **20** (9): 567-587.

8. M. Totrov; *Atomic Property Fields: Generalized 3D Pharmacophoric Potential for Automated Ligand Superposition, Pharmacophore Elucidation and 3D QSAR.* Chem. Biol. Drug Des.; *2008*. **71** (1): 15-27.

9. S.E. O'Brien, D.G. Brown, et al.; *Computational tools for the analysis and visualization of multiple protein–ligand complexes.* Journal of Molecular Graphics and Modelling; *2005*. **24** (3): 186-194.

10. R.A. Powers and B.K. Shoichet; *Structure-Based Approach for Binding Site Identification on AmpC β-Lactamase.* J. Med. Chem.; *2002*. **45** (15): 3222-3234.

11. S.E. Nichols, R.A. Domaoal, et al.; *Discovery of Wild-Type and Y181C Mutant Non-nucleoside HIV-1 Reverse Transcriptase Inhibitors Using Virtual Screening with Multiple Protein Structures.* J. Chem. Inf. Model.; *2009*. **49** (5): 1272-1279.

12. H. Mitsuya, K.J. Weinhold, et al.; *3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro.* Proc. Natl. Acad. Sci.; *1985*. **82** (20): 7096-7100.

13. D.D. Richman; *Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy.* J. Virol.; *1994*. **68** (3): 1660.

14. D.V. Havlir, S. Eastman, et al.; *Nevirapine-resistant human immunodeficiency virus: kinetics of replication and estimated prevalence in untreated patients.* J. Virol.; *1996*. **70** (11): 7894-7899.

15. S. Rhee, W. Fessel, et al.; *HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance.* J. Infect. Dis.; *2005*. **192**: 456 - 465.

16. V.A. Johnson, F. Brun-Vezinet, et al.; *Update of the Drug Resistance Mutations in HIV-1: December 2008.* Topics in HIV medicine : a publication of the International AIDS Society, USA; *2008*. **16** (5): 138-145.

17. Z. Zhou, M. Madrid, et al.; *Effect of a Bound Non-Nucleoside RT Inhibitor on the Dynamics of Wild-Type and Mutant HIV-1 Reverse Transcriptase.* J. Am. Chem. Soc.; *2005*. **127** (49): 17253-17260.

18. M. Madrid, J.A. Lukin, et al.; *Molecular dynamics of HIV-1 reverse transcriptase indicates increased flexibility upon DNA binding.* Proteins: Struct., Funct., Bioinf.; *2001*. **45** (3): 176-182.

19. I. Bahar, B. Erman, et al.; *Collective Motions in HIV-1 Reverse Transcriptase: Examination of Flexibility and Enzyme Function.* Journal of molecular biology; *1999*. **285** (3): 1023-1037.

20. K. Das, S.G. Sarafianos, et al.; *Crystal Structures of Clinically Relevant Lys103Asn/Tyr181Cys Double Mutant HIV-1 Reverse Transcriptase in Complexes with ATP and Non-nucleoside Inhibitor HBY 097.* Journal of molecular biology; *2007*. **365** (1): 77-89.

21. H.A. Carlson; *Protein flexibility and drug design: how to hit a moving target.* Curr. Opin. Chem. Biol.; *2002*. **6** (4): 447-452.

22. H.A. Carlson and J.A. McCammon; *Accommodating Protein Flexibility in Computational Drug Design.* Mol. Pharmacol.; *2000*. **57** (2): 213-218.

23. Y. Hsiou, J. Ding, et al.; *Structure of unliganded HIV-1 reverse transcriptase at 2.7 å resolution: implications of conformational changes for polymerization and inhibition mechanisms.* Structure; *1996*. **4** (7): 853-860.

24. K. Das, J.D. Bauman, et al.; *High-resolution structures of HIV-1 reverse transcriptase/TMC278 complexes: Strategic flexibility explains potency against resistance mutations.* Proc. Natl. Acad. Sci. U. S. A.; *2008*. **105** (5): 1466-1471.

25. F. Rodríguez-Barrios, J. Balzarini, and F. Gago; *The Molecular Basis of Resilience to the Effect of the Lys103Asn Mutation in Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors Studied by Targeted Molecular Dynamics Simulations.* J. Am. Chem. Soc.; *2005*. **127** (20): 7570-7578.

26. G. Maga, M. Radi, et al.; *Discovery of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase Competing with the Nucleotide Substrate.* Angew. Chem.; *2007*. **119** (11): 1842-1845.

27. R. Esnouf, J. Ren, et al.; *Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors.* Nat. Struct. Mol. Biol.; *1995*. **2** (4): 303-308.

28. J. Balzarini; *Current Status of the Non-nucleoside Reverse Transcriptase Inhibitors of Human Immunodeficiency Virus Type 1.* Curr. Top. Med. Chem.; *2004*. **4** (9): 921-944.

29. L. Kohlstaedt, J. Wang, et al.; *Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor.* Science; *1992*. **256** (5065): 1783-1790.

30. S.G. Sarafianos, B. Marchand, et al.; *Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition.* Journal of molecular biology; *2009*. **385** (3): 693-713.

31. J. Mendieta, C.E. Cases-González, et al.; *A Mg2+-induced conformational switch rendering a competent DNA polymerase catalytic complex.* Proteins: Struct., Funct., Bioinf.; *2008*. **71** (2): 565-574.

32. K.A. Paris, O. Haq, et al.; *Conformational Landscape of the Human Immunodeficiency Virus Type 1 Reverse Transcriptase Non-Nucleoside Inhibitor Binding Pocket: Lessons for Inhibitor Design from a Cluster Analysis of Many Crystal Structures.* J. Med. Chem.; *2009*. **52** (20): 6413-6420.

33. J.K. Wegner, H.W.T. Van Vlijmen, and C.W.M. Boutton. *Phenotype prediction method* W.I.P. Organization *2008* WO2008/065180A1

34. Z.K. Sweeney, S.F. Harris, et al.; *Design of Annulated Pyrazoles as Inhibitors of HIV-1 Reverse Transcriptase.* J. Med. Chem.; *2008*. **51** (23): 7449-7458.

35. Z. Wang, B. Wu, et al.; *Synthesis and biological evaluations of sulfanyltriazoles as novel HIV-1 non-nucleoside reverse transcriptase inhibitors.* Bioorg. Med. Chem. Lett.; *2006*. **16** (16): 4174-4177.

36. N.T. Parkin, S. Gupta, et al.; *The K101P and K103R/V179D Mutations in Human Immunodeficiency Virus Type 1 Reverse Transcriptase Confer Resistance to Nonnucleoside Reverse Transcriptase Inhibitors.* Antimicrob. Agents Chemother.; *2006*. **50** (1): 351-354.

37. P. Gerondelis, R.H. Archer, et al.; *The P236L Delavirdine-Resistant Human Immunodeficiency Virus Type 1 Mutant Is Replication Defective and Demonstrates Alterations in both RNA 5'-End- and DNA 3'-End-Directed RNase H Activities.* J. Virol.; *1999*. **73** (7): 5803-5813.

38. *ICM* Molsoft L.L.C., 2009 Version 3.6d.

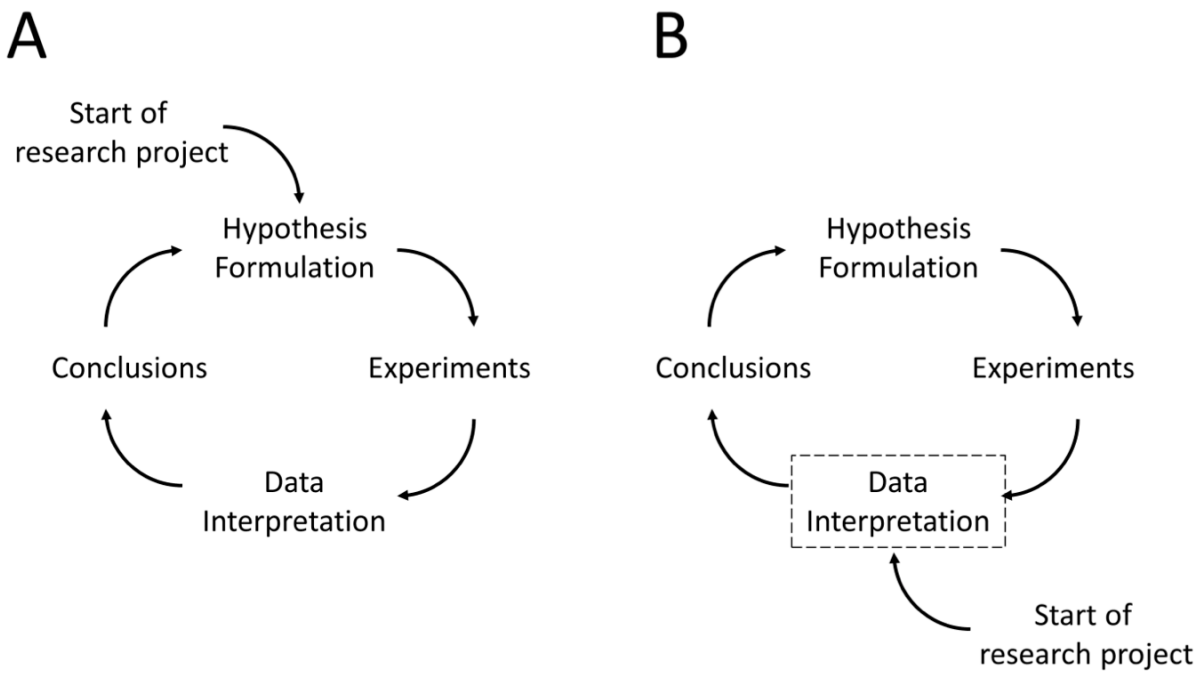
39. *Marvin Beans pKa Prediction tool* ChemAxon 2006 Version 4.1

40. R. Najmanovich, J. Kuttner, et al.; *Side-chain flexibility in proteins upon ligand binding.* Proteins: Struct., Funct., Bioinf.; *2000*. **39** (3): 261-268.

41. *ICM user manual*; *2009*; Molsoft L.L.C.,: La Jolla.

Chapter 8

Conclusions and Future Perspectives



Contents

8.1 Personal observations in computational chemistry.

8.1.1 Primarily a Scientist. After four years of working as a PhD candidate I should like to start this chapter with several personal observations made during that time (these ought to be regarded as just that, personal observations).

Computational disciplines routinely handle datasets of immense size. Intuitive tools, slick graphical user interfaces (GUI) and standardized data formatting rest on advanced computational concepts. However, they also bring about two major pitfalls that should not be overlooked.

Firstly, in the predictable world of computational analysis it is sometimes difficult to remember concepts relevant in laboratory work like experimental error and reproducibility. It is tempting to treat K_i values that are presented as factual data points, however a 'data smear' is a more appropriate description. The trouble is that one doesn't know what the *true* value is for these 'data smears' while these points are often treated as if one does. Success is not always guaranteed when reproducing an experiment and no two experiments are identical. Whereas in computational approaches reproducing an experiment will usually lead to success and two experiments can be run on opposite sides of the world in different labs leading to identical results (down to three decimals or better).

Secondly, when confronted with a large data set it is often difficult to find a good starting location. However, the computational scientist should always remember to carefully curate the data to the best of his ability before embarking on any modeling approach. "Not all data points are created equal", and a model is as good as the error in the data. One should remember that the data feeding computational work is always a hand me down from other scientists. One is not always as lucky as to personally know this previous owner.

Hence, the most important conclusion from my thesis is that, despite the highly organized and reproducible fashion is which computational experiments are performed, computational chemistry remains a true scientific discipline with errors, uncertainty and limited capabilities. However, the location where this uncertainty resides is often not easily spotted in any experiment. Therefore no model or analysis should ever be treated as routine and getting to know one's data always pays off in the long run.

8.2 Observations from this thesis

8.2.1 Meeting the pre-set aims. This thesis focuses on knowledge-based computational approaches to combine data from different disciplines that are relevant in medicinal chemistry and drug discovery. The rationale was that these disciplines, chemistry, biology and bioactivity, are *complementary* and that there is much to be gained by combining them. After several research chapters we can conclude that this is indeed the case. Combining data, the addition of extra information improves models. However, when the matter in which this data is combined is not accurately represented by the descriptors (e.g., combining allosteric and orthosteric compounds without differentiating in the binding site) the effect is that the final combined models are inferior to individual models. Only a thorough validation can spot these problems and care should be taken to validate the performance of a model on each target individually rather than over all targets as this masks a single target that performs worse than the average.

8.2.2 PCM, a technique with many names. Arguably the most important subject of the thesis is proteochemometric (PCM) modeling, which plays a central role in the majority of the chapters and is reviewed in **chapter 2**. We find that PCM is a technique that is gaining ground in scientific literature as it is applied by diverse groups to diverse targets. At the same time the diverse user base also leads to fragmentation in literature as the same technique carries multiple names (PCM,[1, 2] chemogenomics,[3] Protein-Ligand fingerprint,[4] Multi-Assay-Based SAR [5]) all of which combine data from related targets. Yet, the difference between these papers resides in what type of data (e.g. chemical structures linked to assay activity, or information about what interactions between a compound and target are possible linked to a docking score) is combined and what are considered 'related targets'.

Historically, most novel approaches can actually be classified within existing parameters based on the type of description included for the target. PCM is no exception and should therefore be classified as a method to statistically derive a Structure-Activity-Relationship (SAR). The same goes for the chemogenomics work by Jacob *et al*. and the Multi-Assay-Based SAR by Ning *et al.*[3, 5] Other approaches, like Protein-Ligand fingerprint, should be classified as structural methods much like the work presented in **chapter 7**.

What to consider related targets (and hence which targets to dismiss when creating PCM models) is highly dependent on the eventual goal of the model. For example, in a receptor deorphanisation experiment, related targets should mean almost any protein of a superfamily, allowing a full sampling of the target space. Conversely, when viral resistance is modeled, the group of related targets should be much narrower, for instance limited to mutated versions of the main protein of interest. Likewise, the descriptor used in the PCM model should also be suitable for the target under consideration as we show in **chapter 3**.

When these limitations are acknowledged, PCM can be very flexible and applied to almost any group of targets as we have shown in this thesis. Furthermore the technique can be applied both pre-clinically and clinically as we will illustrate below.

8.2.3 Novel pre-clinical applications of PCM. In literature, PCM is mainly applied in a conventional way meaning the modeling of a series of ligands to a series of targets. However, there are new methods of application flowing directly from the potential combination of different target spaces. Some of these new application areas, like concurrent allosteric and orthosteric SAR modeling, are covered hypothetically in **chapter 2** and others are dealt with in the research chapters of this thesis.

For example, in **chapter 4** we use PCM to concurrently model orthologs and paralogs. This combination provides a way to incorporate historic data. The adenosine receptors provide a superb example as early work investigating this receptor family was actually performed on rat receptor orthologs.[6] In modern science it is particularly relevant not to reinvent the wheel. Using computational tools one has the freedom to incorporate high amounts of data to arrive at a preliminary understanding of the ligand – target space *before* any 'wet' experiment is performed. However, a lot of work has been published and forgotten, in particular if that work was done on other species orthologs. Yet, the information contained therein is still very relevant. Addition of historic compounds found to be active can help improve future work by making the model predictions more robust, but also by widening the applicability domain (in this case the chemical space wherein the model can make reliable predictions).

Likewise, we explored the limitations of target space. Theoretically infinite, we found that in practice there are boundaries. These limitations were explored in **chapters 5** and **6**. We found that these limitations can indeed be quantified based on target similarity. This allows a measure of reliability to be added to model predictions. Furthermore, preliminary experiments were started when we applied PCM to create a class A GPCR wide predictive model.

In essence such a model would have the potential to describe the full class A receptor ligand interaction space. However we found that the target definition is still cumbersome and limits the predictive capabilities of such a model. While we were able to train a model on almost all class A GPCRs (limited only by the availability of active compounds on certain subtypes), we found that the variation in binding pockets of class A GPCRs is still significant. Previous papers have been published that were successful in distinguishing between receptor subtypes based on just these binding pockets, more specifically the binding pockets located within the trans-membrane (TM) domains. We could repeat that distinction but this does not guarantee that these residues are the ones important for ligand binding. In other words, it is possible that the actual similarity measurement between the receptors based on these binding pockets was possibly not in line with the similarity as it is in bioactivity space.

Recently a paper was published by Cheng *et al.* where the authors compare a PCM based approach to an approach which relies on the consensus of 100 individual QSAR models (deemed multitarget-QSAR).[7] The authors find the PCM based approach to predict a large number of false positives. In my view the large number of false positives can be attributed to the usage of the binding pocket defined by Gloriam *et al.* In 2010 Wu *et al.* published the crystal structure of the CXCR4 receptor, showing that binding pockets among GPCRs can differ to a large degree.[8] In the case of the CXCR4 receptor the ligand binds much higher (closer to the extracellular side of the receptor) compared to the previously published structures. Hence other residues are involved in this interaction.

Therefore, a possible follow up for this work would be to distinguish GPCRs based on their ligand type (i.e. purine, peptide, etc) and to define a binding pocket for each of these subclasses. Work by Surgand *et al*. can be a good starting point to define binding pockets per receptor cluster.[9] Furthermore, an increase of available GPCR crystal structures can be instrumental herein. PCM can form the bridge from receptors that have an available crystal structure to receptors that have a similar ligand but lack the availability of a crystal structure.

Lastly, PCM can be of instrumental value in optimizing compounds that should have an effect on multiple isoforms of a protein target. An example application is given in **chapter 5**. The target is formed by a viral protein, HIV-1 Reverse Transcriptase, which has been shown to mutate quickly under selective pressure by treatment with anti-retroviral drugs. An ideal drug is active on the isoform that occurs most frequently, wild type, but also on mutants that arise during treatment. When information about frequently occurring mutations is available, the optimal drug candidate can already be selected in the preclinical phase.

8.2.4 Novel clinical applications of PCM. Finally, applications of PCM are not limited to preclinical research. In **chapter 6** we show how PCM can be used to select and optimize treatment regimens for patients infected with HIV. Hence PCM can be a tool to create personalized treatment protocols. In **chapter 6** the chemical space is formed by the drugs that are FDA approved and the target space is formed by the mutants that have been characterized in an assay based personalized medicine methodology that has been approved for clinical use.

In conclusion, PCM can make robust models by using existing chemical, sequence and biological data. Compared to models created from only chemical information or models created with sequence information only, PCM is shown to create more robust models and hence improved predictions.

8.2.5 Linking crystal structures by consensus structures. While we have shown that the combination of multiple disciplines in the form of bioactivity and chemistry can improve predictability of models, more efficient combination of information from a single discipline can also lead to new insights. In **chapter 7** we introduce ‘consensus structures’. The consensus structures reinterpret existing information present in crystal structures, leading to new insights and visualization of hidden information. We demonstrate this by applying the method to a target that has been studied since 1995, HIV Reverse Transcriptase. Yet we were still able to identify novel features of the allosteric non-nucleoside reverse transcriptase binding pocket that had not been previously described.

This method relies on the presence of multiple crystal structures from a single target. This is the case for a number of enzyme targets, such as HIV-RT, but not for the GPCR superfamily yet. GPCRs have notoriously been very difficult to crystalize, moreover the first crystal structures have been published as recently as 2007 and 2008.[10, 11] Interestingly, now, in 2012, at least a dozen different adenosine A_{2A} receptor crystal structures have been reported, also with different ligands. Combined with the exponential growth rate of the number of structures in the PDB, the case of the adenosine receptors demonstrates that it is likely that the consensus structures can prove valuable on other targets in the near future, such as GPCRs.

8.3 General conclusions from the thesis

From the research chapters I conclude that novel approaches to better mine existing data are not definitive solutions. However, these approaches can be seen as complementary to existing methods as they provide answers that cannot be obtained with traditional methods.

Another important conclusion to be drawn is the utmost importance of thorough validation of computational models and predictions. Using a (larger) dataset with more descriptors (increasing feature space) to train models leads to a larger risk for chance correlations at the same time. With methods sensitive to predict a single drug – target interaction, presence of wrongly annotated drug – target interactions in the training set can severely reduce the quality of model predictions (as shown in **chapters 4** and **6**). However, these prediction errors are in the eye of the beholder as the model merely predicts something that is not expected by the scientist but still accurately reflects what is in the training set.

It is possible to estimate the total number of wrongly annotated values and the average resulting error from several samples, subsequently a confidence value can be estimated for each prediction. However, the cause of a model prediction uncertainty (error) can be one of many reasons, and as such it is virtually impossible to compensate for these errors before training a model. For instance a prediction uncertainty can be caused by an error in annotation, which resides in a small fraction of the total data points. Conversely, an uncertainty in predictions can also be due to large error on a single compound, leading to a small average uncertainty for the full set. Thirdly, an uncertainty can also be caused by a structural bias for a certain scaffold (see **chapter 4**). Moreover, a prediction error or 'wrong' can simply be that measured activity values are lower in a certain assay readout compared to another, therefore values obtained using that particular assay could be classified as inactive rather than active as they are lower than a certain threshold.

8.4 Future Perspectives for PCM.

8.4.1 Complementary tool. As concluded above, these novel approaches, like PCM, will never be a replacement for existing, e.g. Quantitative Structure-Activity Relationship (QSAR), models. The technique can rather be seen as complementary to QSAR. QSAR can be of high interest when optimizing for a single target, in this case PCM does not necessarily improve upon QSAR. However, in the early phase hit discovery, PCM can help locate hits for this particular target by making use of its similarities to nearest neighbors. The great strength comes from the fact that PCM is an expansion of current techniques, using ready available data.

As shown in this thesis, PCM is capable of creating a single model that predicts the activity of a single molecule on a large group of targets. Compounds that have a described bioactivity on not one but many targets are known as poly-pharmacologic compounds. Further expanding on methods like PCM might make predictive poly-pharmacology models a reality. I will illustrate this using several preclinical scenarios.

8.4.2 Compounds hitting a number of targets. Small molecules (or peptides) that should hit a number of targets are expected to become more relevant as novel drugs.[12] Single target hitting drugs have been pursued for many years; however it has been shown that many of the successful drugs actually display a poly-pharmacologic profile. Examples include: kinase inhibitors and anti-psychotics, drugs used in the treatment of cancer and depression.[13, 14] However, compounds that are active on multiple mutants of a virology target (HIV, Hepatitis, Influenza) are also polypharmacologic. We have shown in this thesis (**chapter 3** and **5**) that PCM is able to capture such a ligand – target interaction space. Likewise, these concepts can be applied in the discovery of new antibiotics. An optimal candidate should be a compound that is active on multiple mutants of a bacterial target. This is particularly relevant as there is a need for new antibiotics.[15]

8.4.3 Compounds with different functional poly-pharmacological effects. While most of this thesis focuses on activity which was defined as affinity, the effect compounds exert on targets differs. For instance in the field of GPCRs several effects are distinguished including: agonism (a compound activates a receptor and thereby one or more signaling pathways), antagonism (a compound blocks receptor activation and thereby signaling pathways), inverse agonism (a compound inactivates a constitutively active receptor), and these effects can also be partially depending on the level of (inverse) activation achieved in comparison to the natural ligand.[16]

In the field of central nervous system drug development it has been put forward that optimal drugs should be selectively non-selective rather than selective.[17] This selective non-selectivity has been described as 'magic shotgun' rather than 'magic bullet'. Roth *et al.* describe that D2 partial agonists which are full agonists on other receptors might be an interesting lead.[17] The discovery of these magic shotgun drugs is an ideal scenario for PCM where the ability to classify, model and predict these responses would be of great value in preclinical candidate selection.

A similar idea is pursued in the search for partial agonists for HCA2, the nicotinic acid receptor.[18] Partial agonism at the HCA2 receptor might lead to favorable effects in the treatment of atherosclerosis while preventing the side effects caused by full agonists.[19] As this receptor is actually known to have two paralogs (the HCA1 and HCA3 nicotinic acid receptor) and no crystal structure is available, this is another area of interest where PCM could add value over existing approaches.[20]

8.4.4 Drug-Target residence time. Not only the effect compounds exert on a target can differ, the average time they remain bound to a target can also vary; this concept is known as drug-target residence time. It has previously been shown that differences in residence time can explain physiological effects not explained by affinity alone.[21, 22] Quantifying this residence time is known as a Structure-Residence-Time-Relationship (SRTR) and preliminary work has appeared in literature.[23] Like the addition of target information can be used to create better SAR models, the addition of target information might lead to better STRT models.

For example, the presence of certain functional groups on compounds might lead to better residence time on a certain targets but not another, through the addition of target information it might be elucidated what protein properties are responsible for this effect. This knowledge could then be used to optimize compounds active on another target that shares these properties. While this is speculation, there is no theoretical limitation.

8.4.5 Side effect screening of hit compounds. A final future prospect for PCM can be found in a completely different area of expertise. It is generally agreed upon that compounds should be selective for a certain target to reduce the chances of serious adverse effects arising from treatment with that compound. While it is currently impossible to accurately predict the complete pattern of interactions caused by a compound, it is possible to predict interaction with known anti-targets with a certain degree of reliability (e.g hERG, see below). PCM might be a tool in the early phase of drug discovery to expand the number of anti-targets for which interactions can be predicted. Examples include: GPCR-mediated side effects (via chemical similarity and binding pocket similarity) and Kinase inhibiting side effects (again via chemical similarity and binding pocket similarity).

Currently it is infeasible to predict the actual reliability of PCM when applied in side effect prediction. However the same principles apply that were used in **chapter 4**, therefore also here are no theoretical limitations to the application of PCM.

8.4.6 Novel developments in machine learning. PCM can also benefit greatly from the developments in the field of machine learning. Until recently, one was forced, when selecting a learning method, to choose between either high accuracy OR high interpretability (e.g. the choice between 'support vector machines' (SVM) or 'partial least squares' (PLS) in the case of PCM).[24, 25] Current developments allow for the combination of these two abilities in 'Random Forests', as we have shown in **chapter 3** and will further elaborate on below.[26]

8.4.7 Exponential growth of processing power. The exponential growth in computational power and available data that is currently taking place in computational chemistry is expected to continue over the coming decade. During my PhD project I have actually experienced an example of this growth first hand and I want to illustrate this with two examples.

The NNRTI data set used in chapters 3 and 5 was obtained already in November 2008. Early 2009 the first initial Leave-One-Sequence-Out (LOSO) experiments were performed on this data set. Using the Blosum protein descriptor (the largest and most training intensive, see chapter 3) training and validation of 14 LOSO models took approximately ***108*** hours using a standard workstation (Core 2 Duo E4600 2.4 GHz and 4GB memory). The models were built using the 'e1071' SVM package.[27]

In March 2012 the experiment was repeated with the same data set. At that point the models were trained using the 'forest' random forest package on an updated workstation (Core i7 860 2.8 GHz and 16 GB memory).[26] Total training time for the 14 LOSO models using the Blosum descriptor was ***2.5*** hours. The performance of the final models was in fact comparable and this represents a decrease in training time of *98 %*. Moreover, repeating this experiment for all fingerprints tested in **chapter 3** (13 times the dataset size) took a mere ***30*** hours. While this is caused by an increase in computational power and the advent of more efficient algorithms it illustrates how the progress in computational chemistry leads to more efficient data processing.

A similar example can be given for our structure-based approach. The consensus structures that are presented in **chapter 7** were originally trained in January 2007. At that point these density maps were calculated in approximately 15 minutes (0.25 hours) per protein structure binding pocket and 2 minutes (0.03 hours) per ligand structure. Hence, when calculating for 36 structures the total time was **10** hours. In March 2012 this experiment was repeated and applied to the structures available for the Adenosine A_{2A} receptor (then a total of 9 structures). The isolated binding pockets and co-crystalized ligands are comparable to those of the crystal structures in **chapter 7**. The total calculation time was now about 4 minutes for the 9 binding pockets combined (0.44 minutes per binding pocket) and less than 1 minute for the ligands combined. The total calculation time was thus 5 minutes (or **0.08** hours), representing a decrease of *97 %* in calculation time.

From these observations we can draw a number of conclusions other than just the experimental speed up. Firstly a dataset considered to be infeasible to model could be the subject of a standard procedure in a mere 3 years. Secondly, when optimizing a method and one is presented with a trade-off between speed and accuracy, it could very well pay off to optimize for accuracy rather than speed. Speed will catch up over the years, accuracy will not.

8.5 Future perspectives for structure-based methods

8.5.1 Millisecond molecular dynamics. It is in this light that we should consider the future perspectives for the structure-based methods. While the consensus structures provide a very valuable tool for the near future, their use might become obsolete by the advent of millisecond molecular dynamics (MD).[28] Molecular dynamics simulates the actual dynamic forces between individual atoms and is not novel. However these computational intensive calculations used to take days for simulations that only simulate nano- to microseconds of real time. The catalyst to make these simulations feasible on a millisecond scale is a specialized computer system named Anton (after Anthony van Leeuwenhoek).[28] Anton consists of 512 or more nodes (custom hardware subunits) designed to work together efficiently with custom software. Already this purpose-built approach has shown that it is capable of performing molecular dynamics calculations of GPCRs and has been used to retrospectively validate force fields currently in use.[29, 30]

The major advantage of MD over consensus structures is that MD is not bound by states of the protein which can be crystallized like consensus structures are. MD can also model transition states *between* states that can be crystalized. An obvious application area is modeling of extracellular loops (ELs) on GPCRs. These loops display an enormous degree of variation between the different GPCR crystal structures yet have also been shown to be important in ligand binding,[31, 32] and allosteric modulation.[33] Furthermore, MD allows the simulation of water molecules in the structure, the presence of which has previously already been shown to be very important.[34, 35]

The downside of MD is the fact that it is limited by the quality of force fields which are always an approximation whereas crystal structures are experimentally derived results. An extensive validation of these MD approaches is therefore key before their mainstream use and this validation has already begun to appear in literature.[30]

8.6 Drug discovery remains a challenging field

8.6.1 The drug discovery problem. When we compare drug discovery with other applied sciences like engineering or physics the most important difference is that in drug discovery we fail to completely understand the system we work on. An engineer designing a plane can accurately simulate the plane in flight *in silico* and the finally built prototype will behave near identical to those simulations. This omits a great deal of the experimental optimizations needed before the first test flight. While developments have progressed quickly in all fields mentioned in this thesis (chemistry, biology, bioactivity, computational approaches) we are not yet capable to fully understand our target organism 'homo sapiens'. Hence we cannot fully simulate a candidate drug in a human *in silico.*

It should however be noted that the first publication of a computational model simulating *a whole cell* (*Mycoplasma genitalium*) has appeared in literature July 20th 2012.[36] Still, one of the underlying causes preventing the simulation of 'homo sapiens' is the sheer dimensionality of the problem. Moreover, simply optimizing a compound to be a perfect binder on a single target is already a complex problem we do not fully understand.

Take the case of logP, a physicochemical property of compounds that cannot be perfectly predicted. Computational tools provide an estimate based on known observations and novel tools are still appearing in literature,[37-39] indicating that we do not fully understand our problem. Likewise, when presented with a crystal structure of a target and co-crystalized ligand, computational tools fail to present perfect predictions, again indicating a lack of complete understanding of the problem.[40, 41] While this is not novel information, it is relevant to be able to consider the drug discovery problem as outlined below.

8.6.2 Single solution for a complex problem. On the level of a drug being prescribed to a patient, the multi-dimensionality of the problem further explodes. The drug is the single solution to this very complex multi-dimensional problem. A very complex problem indeed as several of the parameters that need to be optimized have contradictory goals. A classic example case is the development of a compound with a target in the central nervous system. It might very well be that this compound requires several hydrogen bonds to have a high activity on a certain protein. However, from literature we known that an increase in hydrogen bond capacity is detrimental for blood-brain-barrier permeability.[42, 43]

Likewise the assumption that nanomolar affinity on a target leads to selectivity for that target is not always true. For instance the discovery of the hERG voltage gated potassium channel as the causative agent of cardiovascular sudden death has caused the withdrawal of several high affinity blockbuster drugs.[44, 45] Unwillingly, in the process of compound binding optimization, one might introduce hERG affinity by optimizing affinity to the actual target. Hence one is simultaneously creating a problem during the process of tackling another. The hERG channel itself is an example of the discovery of novel factors in the multi objective problem that change the drug discovery landscape continuously. Here I will classify these factors as the 'unknown problem'.

8.6.3 The unknown problem. Each drug is unique and each drug discovery track will encounter different hurdles. As we have outlined these hurdles might relate to the nature of the target or they might relate to the nature of the compound under development. Moreover, with the further unraveling of the target organism, possible *novel* problem areas are appearing in literature.[46] It might sound as an impossible task to create a drug, and the decrease in output of novel drugs supports this grim image.[47]

However, computational tools *can* be used to improve success rate, the key is to make use of the data available. Part of the solution to this problem is to be smart rather than to use the brute force method. Already universities have demonstrated incredible hitrates up to 70 % as shown by De Graaf *et al.* or high accuracy in crystal structures prediction Kufareva *et al.*, being universities their budget is limited and they cannot resort to brute force.[48, 49] Novel approaches are required in hit identification and the hit should preferably already meet several of the downstream requirements. [50]

8.6.4 Incorporating computational methods into existing research lines. Computational tools are cheap, quick and have become much more reliable. Can computational methods add value? I would argue that they can add value in any (drug) research project. But the question is how or when these methods add value. Naturally the nature of this added value depends on the specific research project, still there are some globally applicable ways in which to use computational tools.

- Helping to formulate a hypothesis that is partially based on previously available public data. Mining for this data is not necessarily limited to one's own field (for example problems one comes across in PK/PD may very well have been solved in the field of QSAR).
- Iterative data interpretation and comparison to known data might help in guiding a project while underway, and to prevent forgetting early lessons learned.

Both these two phases of a project consist of data interpretation. This is also what is known as e-Science. E-Science approaches can often obtain surprising results from unexpected sources. An example is work by Frijters *et al.*[51] Using only literature mining they identified novel associations between genes, drugs, pathways and diseases that have a high probability of being biologically valid.

In particular the fact that they did this automatically makes this a very useful tool. In this example the integration of computational approaches proved successful. Addition of computational methods into existing projects should not be cumbersome or difficult, it consists of merely shifting the starting point in a research project (**Figure 8.1**). The rationale here is that added data is added value as long as the data consists of information and not noise.

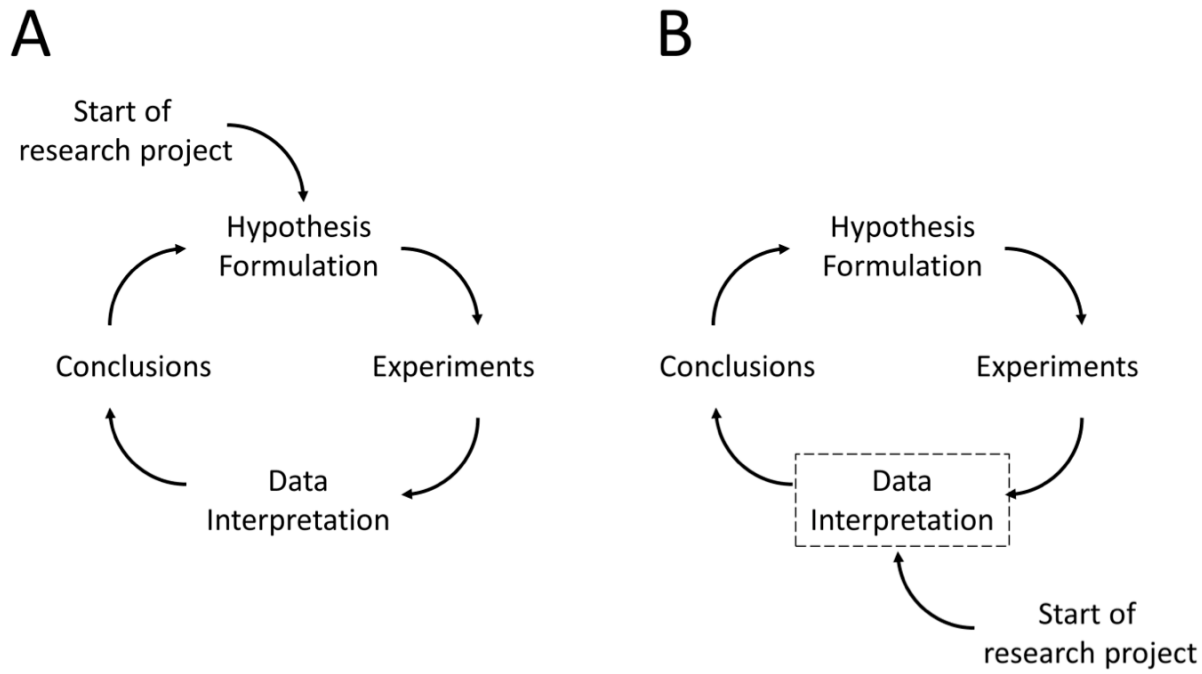


Figure 8.1: How computational methods can be integrated in existing research projects. (A) The classical way the scientific method guides research. (B) Addition of computational approaches. The fundamental set-up does not change, rather computational methods can be introduced early on in a research project (dashed box). By mining public data available these methods can help in constructing a solid hypothesis. Likewise, data gathered from experiments can be interpreted to arrive at a final conclusion.

8.6.5 Public data is not everything. Let us examine the public data we used for our model creation in **chapters 3** and **4** and which we propose to be used for hypothesis formulation in research. More specifically, let us compare the information used by companies to found their research on (proprietary data) with this data which is available in the public domain. While not much has been published, there seems to be a separation between public data and proprietary data (**Figure 8.2**). [52]

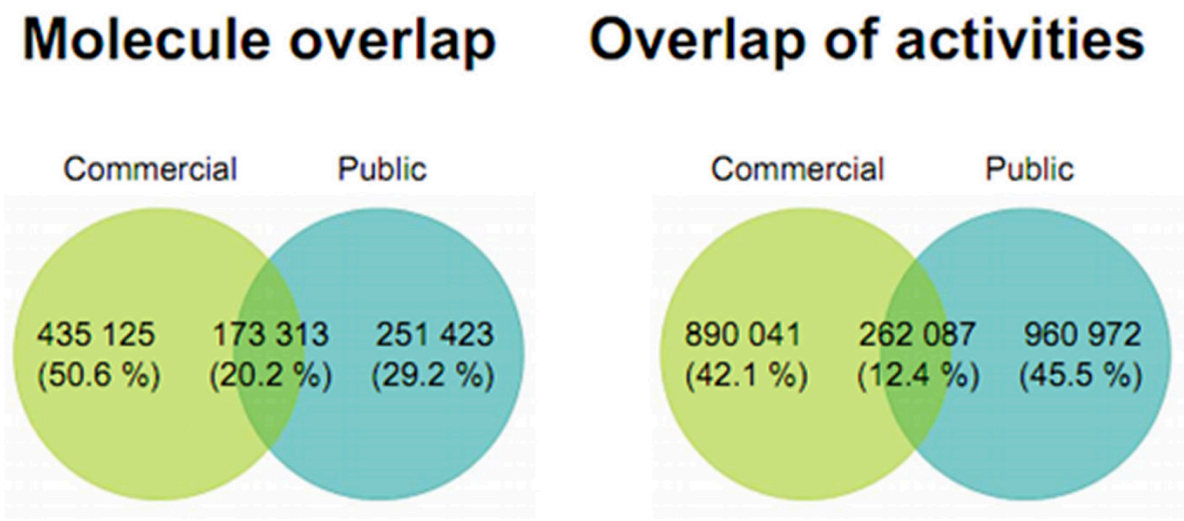


Figure 8.2: Overlap between public and proprietary databases. Numbers are for molecules associated with an activity. An activity is defined as a unique combination of Uniprot ID, small molecule, activity value, activity type and activity relation (Adapted from: Pekka Tiikkainen and Lutz Franke, Analysis of Commercial and Public Bioactivity Databases, 2011 [52]).

Figure 8.2 shows that, even with the quick growth of publicly available data, pharmaceutical companies still own a siginificant part of the structure-activity space that is unavailable in the public domain. Therefore it is elementary that pharmaceutical companies need to combine public and proprietary data and cannot merely rely on their in-house data.

8.7 Final conclusion

At the end of this thesis I should like to draw one final conclusion based on the research chapters covered in this thesis and the thesis itself. This conclusion is that universities and commercial companies should embark on collaborative research efforts. The academic and the commercial mindset are fundamentally different. None can be considered superior by any standard, but fact is that these mindsets are complementary to a large degree. Therefore it stands to reason that, like we show in this thesis, synergistic effects can result from combining these mindsets. With the advent of large academic-commercial cooperation platforms (so-called public-private partnerships, such as TIPharma, the Innovative Medicines Initiative of the EU, and the cooperation that was the source of this thesis) combining these mindsets is exactly what is happening in drug discovery…

8.8 References

1. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
2. J. Wikberg, M. Lapinsh, and P. Prusis; *Proteochemometrics: A tool for modelling the molecular interaction space;* in *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*; H. Kubinyi and G. Müller; Editors. *2004*. p. 289 - 309.
3. L. Jacob, B. Hoffmann, et al.; *Virtual screening of GPCRs: An in silico chemogenomics approach.* BMC Bioinformatics; *2008*. **9** (1): 363-379.
4. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.
5. X. Ning, H. Rangwala, and G. Karypis; *Multi-Assay-Based Structure-Activity Relationship Models: Improving Structure-Activity Relationship Models by Incorporating Activity Information from Related Targets.* J. Chem. Inf. Model.; *2009*. **49** (11): 2444-2456.
6. B.B. Fredholm, A.P. IJzerman, et al.; *International Union of Basic and Clinical Pharmacology. LXXXI. Nomenclature and Classification of Adenosine Receptors—An Update.* Pharmacol. Rev.; *2011*. **63** (1): 1-34.
7. F. Cheng, Y. Zhou, et al.; *Prediction of chemical-protein interactions: multitarget-QSAR versus computational chemogenomic methods.* Mol. BioSyst.; *2012*.
8. B. Wu, E.Y.T. Chien, et al.; *Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists.* Science; *2010*. **330** (6007): 1066-1071.

9. J.-S. Surgand, J. Rodrigo, et al.; *A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **62** (2): 509-538.
10. V.P. Jaakola, M.T. Griffith, et al.; *The 2.6 Angstrom Crystal Structure of a Human A2A Adenosine Receptor Bound to an Antagonist.* Science; *2008*. **322** (5905): 1211-1217.
11. S.G.F. Rasmussen, H.-J. Choi, et al.; *Crystal structure of the human [bgr]2 adrenergic G-protein-coupled receptor.* Nature; *2007*. **450** (7168): 383-387.
12. A.D. Boran and R. Iyengar; *Systems approaches to polypharmacology and drug discovery.* Curr. Opin. Drug Discovery Dev.; *2010*. **13** (3): 297-309.
13. A.L. Hopkins, J.S. Mason, and J.P. Overington; *Can we rationally design promiscuous drugs?* Curr. Opin. Struct. Biol.; *2006*. **16** (1): 127-136.
14. G.R. Zimmermann, J. Lehár, and C.T. Keith; *Multi-target therapeutics: when the whole is greater than the sum of the parts.* Drug Discov. Today; *2007*. **12** (1–2): 34-42.
15. World Health Organisation. *Antimicrobial resistance*. 2012 [cited 2012 March 16]; Available from: http://www.who.int/mediacentre/factsheets/fs194/en/.
16. K. Terry; *Principles: Receptor theory in pharmacology.* Trends Pharmacol. Sci.; *2004*. **25** (4): 186-192.
17. B.L. Roth, D.J. Sheffler, and W.K. Kroeze; *Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia.* Nat Rev Drug Discov; *2004*. **3** (4): 353-359.
18. W. Soudijn, I. van Wijngaarden, and A.P. IJzerman; *Nicotinic acid receptor subtypes and their ligands.* Medicinal Research Reviews; *2007*. **27** (3): 417-433.
19. T. van Herk, J. Brussee, et al.; *Pyrazole Derivatives as Partial Agonists for the Nicotinic Acid Receptor.* J. Med. Chem.; *2003*. **46** (18): 3945-3951.
20. S. Offermanns, S.L. Colletti, et al.; *International Union of Basic and Clinical Pharmacology. LXXXII: Nomenclature and Classification of Hydroxy-carboxylic Acid Receptors (GPR81, GPR109A, and GPR109B).* Pharmacol. Rev.; *2011*. **63** (2): 269-290.
21. R.A. Copeland, D.L. Pompliano, and T.D. Meek; *Drug-target residence time and its implications for lead optimization.* Nat. Rev. Drug Discovery; *2006*. **5** (9): 730-739.
22. D. Swinney; *The role of binding kinetics in therapeutically useful drug action.* Curr. Opin. Drug Discovery Dev.; *2009*. **12** (1): 31-39.
23. G. Tresadern, J.M. Bartolome, et al.; *Molecular properties affecting fast dissociation from the D2 receptor.* Bioorg. Med. Chem.; *2011*. **19** (7): 2231-2241.
24. C. Cortes and V. Vapnik; *Support-vector networks.* Machine Learning; *1995*. **20** (3): 273-297.

25. S. Wold, A. Ruhe, et al.; *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses.* SIAM journal on scientific and statistical computing; *1984*. **5** (3): 735.

26. A. Liaw and M. Wiener; *Classification and Regression by randomForest.* R News; *2002*. **2** (3): 18-22.

27. E. Dimitriadou, K. Hornik, et al. *Misc Functions of the Department of Statistics (e1071)* TU Wien 2006 1.5-15

28. D.E. Shaw, R.O. Dror, et al.; *Millisecond-scale molecular dynamics simulations on Anton*; in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis2009*; ACM: Portland, Oregon. 1-11.

29. A.C. Kruse, J. Hu, et al.; *Structure and dynamics of the M3 muscarinic acetylcholine receptor.* Nature; *2012*. **482** (7386): 552-556.

30. K. Lindorff-Larsen, P. Maragakis, et al.; *Systematic Validation of Protein Force Fields against Experimental Data.* PLoS One; *2012*. **7** (2): e32131.

31. M.C. Peeters, G.J.P. Van Westen, et al.; *Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation.* Trends Pharmacol. Sci.; *2011*. **32** (1): 35-42.

32. M.C. Peeters, G.J.P. Van Westen, et al.; *GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor.* The FASEB Journal; *2011*. **25** (2): 632-643.

33. M.C. Peeters, L.E. Wisse, et al.; *The role of the second and third extracellular loops of the adenosine A1 receptor in activation and allosteric modulation.* Biochemical Pharmacology; *2012*. **84** (1): 76-87.

34. V. Katritch, V.-P. Jaakola, et al.; *Structure-Based Discovery of Novel Chemotypes for Adenosine A2A Receptor Antagonists.* J. Med. Chem.; *2010*. **53** (4): 1799-1809.

35. W. Liu, E. Chun, et al.; *Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions.* Science; *2012*. **337** (6091): 232-236.

36. Jonathan R. Karr, Jayodita C. Sanghvi, et al.; *A Whole-Cell Computational Model Predicts Phenotype from Genotype.* Cell; *2012*. **150** (2): 389-401.

37. A.K. Ghose, V.N. Viswanadhan, and J.J. Wendoloski; *Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods.* J. Phys. Chem.; *1998*. **102** (21): 3762-3772.

38. C. Kramer, B. Beck, and T. Clark; *A Surface-Integral Model for Log POW.* J. Chem. Inf. Model.; *2010*. **50** (3): 429-436.

39. L. Xing and R.C. Glen; *Novel Methods for the Prediction of logP, pKa, and logD.* J. Chem. Inf. Comput. Sci.; *2002*. **42** (4): 796-805.

40. E. Kellenberger, J. Rodrigo, et al.; *Comparative evaluation of eight docking tools for docking and virtual screening accuracy.* Proteins: Struct., Funct., Bioinf.; *2004*. **57** (2): 225-242.

41. J.B. Cross, D.C. Thompson, et al.; *Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy.* J. Chem. Inf. Model.; *2009*. **49** (6): 1455-1474.

42. F. Ooms, P. Weber, et al.; *A simple model to predict blood–brain barrier permeation from 3D molecular fields.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease; *2002*. **1587** (2–3): 118-125.

43. P. Crivori, G. Cruciani, et al.; *Predicting Blood−Brain Barrier Permeation from Three-Dimensional Molecular Structure.* J. Med. Chem.; *2000*. **43** (11): 2204-2216.

44. M.C. Sanguinetti and M. Tristani-Firouzi; *hERG potassium channels and cardiac arrhythmia.* Nature; *2006*. **440** (7083): 463-469.

45. G.-N. Tseng; *IKr: The hERG Channel.* Journal of Molecular and Cellular Cardiology; *2001*. **33** (5): 835-849.

46. J. Dudley, E. Schadt, et al.; *Drug Discovery in a Multidimensional World: Systems, Patterns, and Networks.* Journal of Cardiovascular Translational Research; *2010*. **3** (5): 438-447.

47. F. Pammolli, L. Magazzini, and M. Riccaboni; *The productivity crisis in pharmaceutical R&D.* Nat Rev Drug Discov; *2011*. **10** (6): 428-438.

48. C. de Graaf, A.J. Kooistra, et al.; *Crystal Structure-Based Virtual Screening for Fragment-like Ligands of the Human Histamine H1 Receptor.* J. Med. Chem.; *2011*. **54** (23): 8195-8206.

49. I. Kufareva, M. Rueda, et al.; *Status of GPCR Modeling and Docking as Reflected by Community-wide GPCR Dock 2010 Assessment.* Structure (London, England : 1993); *2011*. **19** (8): 1108-1126.

50. A. Bender, D. Bojanic, et al.; *Which aspects of HTS are empirically correlated with downstream success?* Curr. Opin. Drug Discovery Dev.; *2008*. **11** (3): 327-337.

51. R. Frijters, M. van Vugt, et al.; *Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases.* PLoS Comput. Biol.; *2010*. **6** (9): e1000943.

52. P. Tiikkainen and L. Franke; *Analysis of Commercial and Public Bioactivity Databases.* J. Chem. Inf. Model.; *2011*. **52** (2): 319-326.

Summary

This thesis focussed on the hypothesis that the combination of data from different research disciplines (here chemistry, biology and bioactivity) will have synergistic effects over methods focussing on a single discipline. We investigated this using several preclinical studies and one study using clinical data.

Chapter 1 introduced and defined common concepts from the world of computational chemistry (including: chemical space, chemical similarity, target space and target similarity). Furthermore the chapter highlights how '-informatics' based approaches have revolutionized the world of chemistry and biology. Similarly it contains a short introduction in structural methods, including the concept of X-ray crystallography. The chapter is concluded by a description of limitations in current methods and sketches why a need for novel approaches, like proteochemometrics, exists.

In **chapter 2** a review of the field of proteochemometrics was provided. We have provided a comprehensive overview of the concept and the full body of work using proteochemometric modelling until 2010 is shown in the primary table. This includes the data set modelled, the descriptors used and the machine learning technique applied. The chapter is concluded with a collection of possible pitfalls and a short outlook on novel machine learning approaches and application methods of proteochemometrics.

In **chapter 3** we introduced five novel descriptors to quantify target similarity and performed an extensive study of these and previously published amino acid descriptors. Amino acid similarity is quantified in a multi dimensional space where distance between amino acids correlates directly with similarity. Hence aromatic amino acids tend to cluster together as do charged amino acids. However this space is the result from a dimensionality reduction applied to a large input matrix and hence forms an approximation of the original input matrix. Therefore it was unknown which method would lead to a descriptor that performs optimal in proteochemometric modelling.

We concluded that the different descriptors all perform similar overall but large differences can occur for individual targets. Hence it is wise to sample different descriptors before embarking on final model training to achieve optimal performance. However we also observed that the inclusion of more information from the original input matrix affected performance of the descriptors in a negative way.

This is likely due to the nature of the data reduction approaches where the first factors tend to explain the majority of the variation in the dataset with subsequent factors explaining iteratively smaller fractions.

In **chapter 4** we performed a preclinical study with the goal to identify novel ligands for the human adenosine receptors. To obtain this goal we wanted to make optimal use of all data available to us in the public domain and combined experimental data obtained in bioassays incorporating human receptors with experimental data obtained on rat receptors. The combination of human and rat data resulted in a larger chemical space and hence our hypothesis was that this would translate to our model being able to identify novel ligands rather than analogues of existing compounds.

Of 54 compounds purchased, six novel high affinity adenosine receptor ligands were confirmed experimentally, one of which displayed an affinity of 7 nM on the human adenosine A_1 receptor. We concluded that our models perform better than current structure-activity modeling, as they were able to retrieve novel ligands (a low average tanimoto similarity to the training set) with a high hit rate (11 %)).

Another preclinical study forms the foundation for **chapter 5**, however here we targeted the application of proteochemometrics in a lead optimization project. Our models were trained on a data set, which was near complete (64 % of the possible compound – target interaction pairs had a pEC_{50} value). Through our model we could complete the missing 36 % with an accuracy that approached the assay accuracy, as we confirmed by a prospective experimental validation.

The high quality of this dataset allowed us to predict bioactivity spectra. In an antiviral drug discovery program, as modelled here, this allows for the selection of a compound that is not only active on the most frequently occurring mutant, but also on the majority of the other mutants. Hence the major contribution of our approach is that the optimal candidate can be selected without the need to perform all required 6,314 experiments. Finally, we demonstrated that we were able to define a model applicability domain based on target similarity.

In **chapter 6** we moved from preclinical studies to a clinical scenario. Here we used the largest dataset in literature to date that was subjected to proteochemometric modeling. The dataset consisted of thousands of unique HIV mutants (both protease and reverse transcriptase were included) on the target side and all clinically available drugs for these targets on the ligand side.

Our results demonstrated that we were able to create models capable of predicting a drug regimen for individual patients. Secondly we showed that PCM models performed better that sequence based approaches. Moreover, we demonstrated that our models are able to capture underlying relationships in the ligand – target space as we could predict the affinity of drugs on mutants not previously encountered. Additionally we could even predict the affinity of drugs on mixtures consisting of multiple unique mutants. Finally we confirmed the ability to define an applicability domain that can determine a reliability measure for each model prediction, an important feature in clinically applied models.

In **chapter 7** we applied a structure based approach rather than a machine learning method. We explored the use of novel techniques to mine the increasing amount of crystal structures available in the protein database. Using HIV reverse transcriptase as a case study we reached new insights on the dynamics of this protein and the changes induced by ligands binding to this target. Furthermore, using a three dimensional density based method deemed 'consensus structures' we identified novel features in a binding pocket that has been extensively studied since 1995. These features, currently unexploited by ligands, will improve the ability to inhibit HIV reverse transcriptase even in the presence of mutations.

Finally, in **chapter 8** we have drawn general conclusions from the thesis and proposed some future perspectives. The main hypothesis of this thesis was that linking information obtained from different disciplines (chemistry, biology, bioactivity) by computational approaches is synergistic. We have shown this to be true in the research chapters. However, the tools we used provided a framework, which relies on data from these disciplines to make predictions. Hence novel insights will lead to the ability to make novel predictions. An example is the case of drug target residence time. This concept is now actively being investigated and the results from these research programs will provide the foothold for the creation of possible residence time predicting models.

Furthermore, we proposed that any active drug discovery or research program should include a preliminary phase where all relevant data is gathered (even from related disciplines). Drawing conclusions in an organized fashion about chemistry active on related targets (which might be distantly related in the case of receptor deorphanization) or about the most efficient way to tune modeling parameters will improve the design of experiments. Hence a knowledge-based preliminary phase will help minimize the costs and time involved in the start of novel research projects.

Samenvatting

De hoofdvraag in dit proefschrift was of het combineren van data uit meerdere disciplines (hier chemie, biologie en bioactiviteit) synergistische effecten zou laten zien boven methoden die slechts data uit één discipline gebruiken. Deze vraag probeerden wij te beantwoorden door middel van een aantal preklinische studies en een studie op een klinische data set.

In **hoofdstuk 1** werden een aantal gebruikelijke concepten uit de wereld van Computational Chemistry geintrodcuceerd (waaronder: chemical space, chemical similarity, target space en target similarity). Tevens werd in dit hoofdstuk kort ingegaan op de manier waarop de '-ínformatics' methoden het chemisch en biologisch onderzoek blijvend hebben veranderd. Daarnaast werd een korte introductie gegeven in Röntgen diffractie kristallografie. Het hoofdstuk werd afgesloten met een opsomming van een aantal tekortkomingen van de huidige methoden waarbij tevens uiteen werd gezet waarom er een behoefte bestaat aan nieuwe methoden zoals proteochemometrics.

Aansluitend werd in **hoofdstuk 2** een review gegeven van proteochemometrics. Het concept werd uitgelegd en het review bevat een nagenoeg compleet overzicht van al het werk uit het veld tot 2010 in een tabel. Tevens bevat deze tabel een overzicht van de data sets waarop PCM werd toegepast, de gebruikte descriptoren en machine learning technieken. Het hoofdstuk werd afgesloten met een opsomming van mogelijke valkuilen en daarnaast nieuwe mogelijkheden en toepassingsgebieden.

In **hoofdstuk 3** werden een 5-tal nieuwe eiwit descriptoren geintroduceerd. Deze en andere eerder gepubliceerde descriptoren (totaal 13) werden aan verschillende analyses onderworpen. Het is in PCM gebruikelijk de overeenkomsten tussen aminozuren te quantificeren door middel van fysisch-chemische eigenschappen. Als gevolg daarvan zullen in een overzicht de aromatische aminzuren een cluster vormen evenals de geladen aminozuren enzovoorts. Om een dergelijk gesimplificeerd beeld te krijgen is het echter noodzakelijk de data te comprimeren. Dientengevolge is de uiteindelijke matrix slechts een benadering van de werkelijkheid en ligt het voor de hand uit te zoeken welke descriptor uiteindelijk het beste werkt in combinatie met PCM.

Wij concludeerden echter dat de descriptoren gemiddeld nagenoeg gelijk presteren desalniettemin kunnen er grote verschillen per eiwit ontstaan. Hierom is het aan te raden voor elk experiment de verschillende descriptoren te testen. Wij observeerden echter ook dat het includeren van meer informatie in de datacompressie niet noodzakelijkerwijs tot betere prestaties leidt.

Het is waarschijnlijk dat dit een direct gevolg is van het feit dat datacompressie methoden het merendeel van de variatie in de eerste factoren verklaren en de hierop volgende factoren steeds minder verklaren.

Hoofdstuk 4 bevatte de eerste experimentele toepassing van PCM, in dit hoofdstuk voerden wij een preklinische studie uit om nieuwe liganden voor de humane adenosine receptoren te identificeren. Om dit doel te bereiken combineerden wij de historische data beschikbaar voor liganden die op humane receptoren getest waren met de data over liganden die op rat receptoren getest waren. Deze combinatie leverde een grotere variatie op in de chemische structuren waartoe ons model toegang had. Onze hypothese was dan ook dat wij met dit model nieuwe liganden konden identificeren in plaats van liganden die minimaal van de bestaande verschillen.

Uiteindelijk kochten wij 54 liganden en valideerden de voorspellingen van ons model experimenteel. Wij identificeerden 6 nieuwe liganden (11 % hit rate) waarvan één een affiniteit had van 7 nM voor de humane A1 receptor. Wij concludeerden dat ons model beter presteert dan de huidige methoden gezien wij een hoge hitrate hadden en nieuwe liganden konden identificeren.

De tweede preklinische studie werd beschreven in **hoofdstuk 5.** Hier werd echter gericht op een latere fase in het medicijn onderzoek (lead optimization). Onze modellen werden getraind op een nagenoeg complete dataset (voor 64 % van de mogelijke ligand – eiwit paren hadden wij een acitiviteitswaarde (pEC_{50})). Door middel van het model konden wij de missende 36 % invullen met een nauwkeirigheid die vergelijkbaar was met die van de experimentele assay.

Vanwege deze hoge nauwkeurigheid konden wij bioactiviteitsspectra voorspellen (waarmee de activiteitsverschillen van liganden op virale mutanten voorspeld kon worden). De belangrijkste ontdekking van dit hoofdstuk is dan ook dat op deze manier het optimale kandidaat medicijn gekozen kan worden zonder dat alle 6,314 experimenten gedaan hoeven te worden. Tot slot voegden we voor dit model een betrouwbaarheidswaarde toe gebaseerd op de aminozuur overeenkomst tussen de mutanten.

Hoofstuk 6 bevatte een studie die gezet was in een klinisch scenario. In dit hoofdstuk werd de grootste dataset waarop ooit PCM uitgevoerd is, gebruikt om persoonlijke behandelplannen voor HIV patienten te voorspellen. De dataset zelf bevatte duizenden unieke HIV mutanten (protease en reverse transcriptase) en alle klinisch beschikbare medicijnen voor de behandeling van HIV.

Onze resultaten demonstreerden dat wij dit inderdaad succesvol kunnen en dat de gebruikte PCM aanpak beter presteert dan de huidige modellen in de kliniek die gebaseerd zijn op alleen de mutanten. Tevens konden wij aantonen dat onze modellen de onderliggende structuur activiteitsrelatie modelleren aangezien de modellen ook succesvol de activiteit van medicijnen op mutanten die niet in de dataset zaten kon voorspellen. Daarnaast konden wij ook de activiteit van medicijnen voorspellen in behandeling van HIV van patienten waarbij niet één enkele mutant dominant was maar meerdere. Tot slot konden wij ook hier een betrouwbaarheid aan de voorspellingen geven, een belangrijk gegeven in een klinische toepassing.

Hoofdstuk 7 is het enige hoofdstuk met een meer structuur georienteerde aanpak. Hier introduceren wij nieuwe technieken die gebruik maken van meerdere kristal structuren van een enkel eiwit om tot nieuwe inzichten te komen, welke met het gebruik van één eiwit niet te breiken zijn. Wij gebruikten HIV reverse transcriptase als model eiwit om tot nieuwe inzichten te komen op het gebied van de interactie tussen dit eiwit en zijn liganden. We konden ook succesvol nieuwe aanknopingspunten in de bindingsplaats identificeren door middel van onze methode ‘consensus structures’ welke nog niet geidentificeerd waren (terwijl er voor dit eiwit sinds 1995 kristal structuren beschikbaar zijn). Het exploiteren van deze aanknopingspunten zal leiden tot betere medicijnen voor het inhiberen van HIV reverse transcriptase.

Tot slot trokken wij in **hoofdstuk 8** een eindconclusie en introduceerden we een aantal toekomst perspectieven. De hoofdstelling van dit proefschrift was dat het samenvoegen van data uit verschillende disciplines (hier chemie, biologie en bioactiviteit) synergistisch is, wat wij ook hebben aangetoond. Echter de methoden die wij hiervoor gebruikten verschaffen een raamwerk, wat afhankelijk is van de bovengenoemde data voor het doen van voorspellingen. Dientengevolge kunnen deze methoden ook in nieuwe gebieden aangewend worden, zoals het voorbeeld van ‘drug target residence time’. Dit concept wordt momenteel uitgebreid onderzocht en deze data kunnen de fundering vormen van modellen die wellicht de ‘residence time’ van nieuwe liganden op een eiwit kunnen voorspellen.

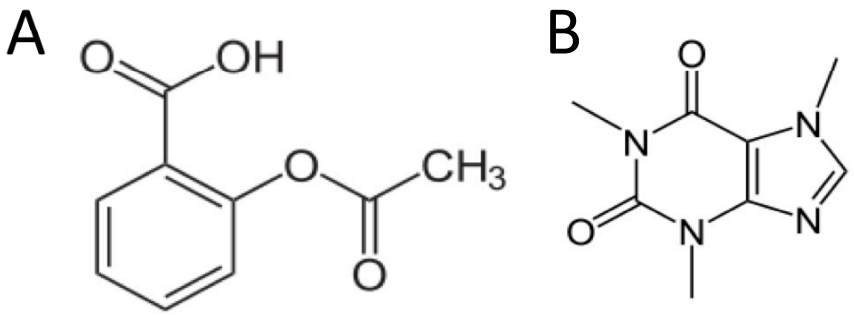
Als laatste stellen wij voor dat elk onderzoeksprogramma uitgebreid moet worden met een preliminaire fase waarin een gedegen literatuuronderzoek alle relevante resultaten uit de literatuur combineert voor een effectieve hypothese vorming. Het verschil met huidige methoden is echter dat men zich hierbij niet moet beperken tot het eigen onderzoeksveld maar ook naar andere disciplines moet kijken. Een voorbeeld is het includeren van liganden voor een eiwit dat slechts weinig gemeen heeft met het eiwit dat onderzocht wordt maar desalniettemin nuttige informatie kan verschaffen. Daarnaast kan het erg lonend zijn om *voor* het maken van een structuur – activiteits model in informatica op zoek te gaan naar nieuwe (mogelijk efficientere) algoritmen. Een uitgebreid onderzoek kan op deze manier een nieuw project een vliegende start verschaffen.

Samenvatting voor leken

De hoofdvraag in dit proefschrift draait om het combineren van data uit gerelateerde disciplines (chemie, biologie en bio-activiteit). Deze concepten verdienen een nadere toelichting, wat wordt ermee bedoeld?

Chemie

Met chemie wordt hier een subklasse van de scheikunde bedoeld, namelijk deze die zich toelegt op de zogenaamde kleine moleculen ('small molecules') die een potentieel effect in het menselijk lichaam hebben. Kleine moleculen betekent precies dat wat men zou verwachten, relatief kleine chemische stoffen (zodat deze beter opgenomen kunnen worden wanneer ze in de vorm van een pil toegediend worden). De kleine moleculen vormen dan ook een van de speerpunten van de moderne medicinale chemie. Voorbeelden hiervan zijn aspirine en caffeine (**Figuur A1**). De gelijksoortigheid van moleculen kan bij medicinale chemie een richtlijn vormen, wanneer een klein molecuul (molecule 1) een bepaald effect heeft is het waarschijnlijk dat een klein molecuul (molecule 2) wat hier sterk op lijkt (gelijksoortig is) een vergelijkbaar effect heeft ('molecular similarity principle').



Figuur A1: De structuur van aspirine (a) en caffeine (b). Beide behoren tot de klasse van kleine moleculen en beinvloeden het menselijk lichaam. Hiermee vallen deze stoffen in de klasse van zogenaamde bio-actieve stoffen.

Biologie

Met het concept biologie wordt in dit proefschrift verwezen naar de biologische aangrijpingspunten van kleine moleculen in het menselijk lichaam ('targets'). Deze aangrijpingspunten zijn vaak eiwitten waarvan de normale signaalverwerking in een ziekteproces verstoord is. Herstel van dit signaalsverwerkingsproces kan in theorie leiden tot genezing en in het optimale geval wordt dit mogelijk gemaakt door small molecules welke in pilvorm toegediend kunnen worden.

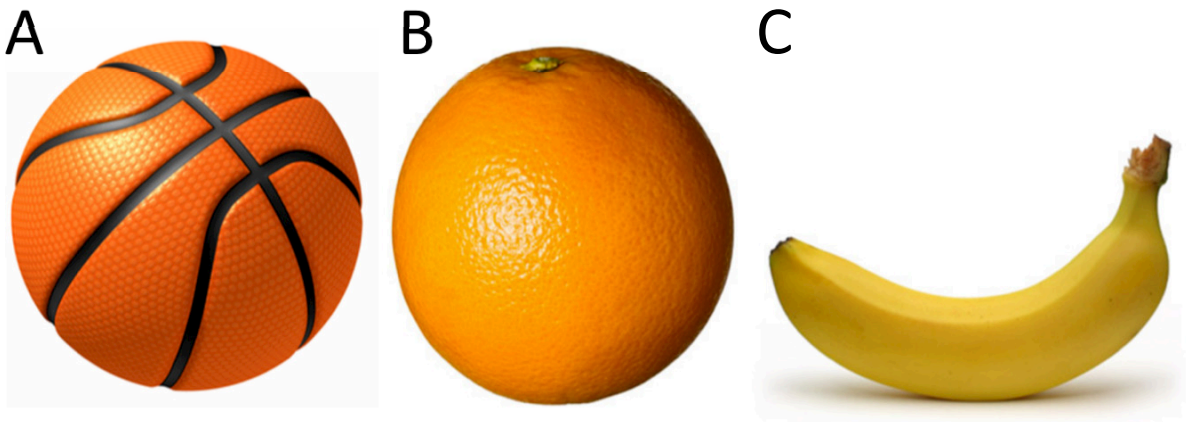
In het geval van targets kan gelijksoortigheid ook gebruikt worden als richtlijn. Stel dat een bepaald target (target A) een verstoorde signaalverweking heeft en er kleine moleculen bekend zijn die de functie herstellen. Het is dan waarschijnlijk dat een target wat hier sterk op lijkt (target B) ook beinvloed zal worden door dezelfde kleine moleculen. Deze effecten staan bekend als bio-activiteit.

Bioactiviteit

De manier waarop en de mate waarin een klein molecuul een target kan beinvloeden wordt bioactiviteit genoemd. Bioactiviteit is een erg breed concept wat strekt van directe effecten tot indirecte effecten. Bioactiviteit kan zelfs resulteren in effecten op het nageslacht. In dit proefschrift wordt de definitie beperkt tot de affiniteit of directe effecten van één or meerdere moleculen op één of meerdere specifieke targets.

Molecular Similarity Principle

De bovengenoemde kleine moleculen (chemie) en targets (eiwitten) kunnen beschreven worden op verscheidene manieren. Gedacht kan worden aan het gewicht van een molecuul op atomaire schaal, hoeveel van een molecuul op te lossen is in 1 liter water etc. Deze eigenschappen waarmee een object te beschrijven is worden ook wel 'Descriptoren' genoemd. De gedachte is dat moleculen die veel op elkaar lijken in deze eigenschappen, ook veel op elkaar zullen lijken in hun effecten op het lichaam. Zie voor een voorbeeld **Figuur A2**, op het eerste gezicht zullen de basketbal en sinaasappel het meeste op elkaar lijken. Echter wanneer de vraag is welk object voedsel is, blijkt dat de sinaasappel en de banaan meer gemeen hebben. De eigenschap waarvan getracht wordt deze met een model te beschrijven (hier of een object voedsel is) aan de hand van de bekende descriptoren, staat bekend als de 'Output Variable'.

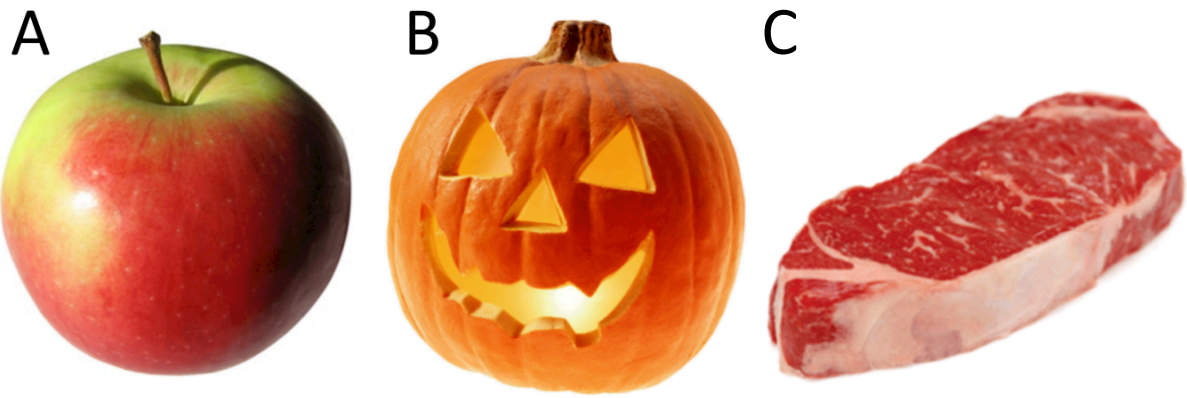


Figuur A2: Het concept van gelijksoortigheid kan veraderlijk zijn en is afhankelijk van de situatie. Hoewel (a) en (b) wellicht het meest gelijkend zijn op het eerste gezicht zal de situatie veranderen wanneer de vraagstelling is welk object waarschijnlijk voedsel is.

Modelleren van bioactiviteit

Zoals gezegd worden in dit proefschrift datasets gebruikt welke chemie, biologie en bio-activiteit combineren. Deze gecombineerde data sets worden vervolgens gebruikt om modellen te creeeren welke voorspellingen kunnen doen over de bio-acitiviteit van (nog) onbekende kleine moleculen. De verwachting is dat de combinatie van deze gegevens zal leiden tot kwalitatief betere voorspellingen dan de voorspellingen van modellen welke gebaseerd zijn op de data van slechts één van deze bovenstaande disciplines. De modellen worden gecreëerd uit sets van moleculen waarvan bekend is welke effecten zij hebben op het menselijk lichaam. Het resulterende model kan aan de hand van de eigenschappen van deze moleculen en van andere, vergelijkbare, moleculen de effecten op het menselijk lichaam voorspellen.

Het concept zal geïllustreerd worden aan de hand van een voorbeeld. De beschreven eigenschap van ons model (output variable) zal zijn om te voorspellen of een object voedsel is. Het model zal gebaseerd worden op de drie objecten uit **Figuur A2**. De descriptoren die gebruikt worden zijn vorm, suikerconcentratie, kleur en de aanwezigheid van rubber. Het resulterende model wordt vervolgens gebruikt om te voorspellen of de objecten uit **Figuur A3** voedsel zijn. Het model zal vervolgens hierbij voor de eerste twee een positief oordeel geven en voor het laatste een negatief oordeel.



Figuur A3: Een aantal onbekende situaties voor ons model. Wanneer de vraagstelling is welk object opgegeten kan worden zal een model getraind op de objecten uit Figuur A2 deze voor (a) met yes (‘active’) beantwoorden, ook wel ‘True Positive’. Een voorspelling voor (b) zou ook active zijn maar een ‘False Positive’. Tot slot kan het model voor (c) voorspellen dat dit object geen voedsel is, ‘False Negative’ genoemd.

De appel is voedsel, het model heeft het hier dus bij het rechte eind ('True Positive'), de pompoen is ook voedsel maar is op dit moment niet eetbaar gezien er een kaars in brandt (vals positief of 'False Positive'). De entrecote wordt ten onrechte beoordeeld als geen voedsel (vals negatief of 'False Negative'). Deze twee foute voorspellingen zijn niet zozeer aan het model te wijten. De false positive wordt veroorzaakt door een tekortkoming in de beschrijving van de objecten (het model heeft nooit geleerd wat een kaars is daar dit niet meegenomen werd in de descriptoren).

Het geval van de false negative kan waarschijnlijk geweten worden aan het feit dat het model slechts geleerd heeft om voor fruit de eetbaarheid te voorspellen, vlees heeft het nooit gezien en dientengevolge kan het hierover geen betrouwbare voorspelling doen. Het is aan de wetenschapper om deze tekortkomingen te voorkomen door de juiste descriptoren te kiezen en het model te baseren op een zo volledig en representatief beeld van de werkelijkheid.

Samenvatting van de hoofdstukken in dit proefschrift

In het **eerste hoofdstuk** worden de concepten en definities omschreven zoals ze in dit proefschrift gebruikt worden. Het **tweede hoofdstuk** bevat een literatuuronderzoek van de gebruikte techniek ('Proteochemometric modeling'). **Hoofdstuk** 3 vergelijkt een aantal descriptoren welke gebruikt worden om meerdere chemische datasets te koppelen. **Hoofdstuk 4** bevat een studie naar modellen welke de adenosine receptoren beschrijven (ook wel verantwoordelijk voor de effecten van caffeine). In **hoofdstuk 5** wordt de techniek toegepast op kandidaat medicijnen met een potentieel HIV remmend effect. **Hoofdstuk 6** gaat hierin een stap verder en bevat modellen welke een persoonlijk behandelingsschema voor HIV patienten kunnen voorspellen. Een vergelijkbare aanpak op moleculaire structuren wordt in **hoofdstuk 7** gepresenteerd. Tot slot bevat **hoofdstuk 8** conclusies en een aantal toekomstperspectieven.

List of publications

G.J.P. Van Westen, A. Hendriks, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Personalized HIV Treatment Regimen Prediction Employing Proteochemometric Models Generated From Antivirogram Data.* **Submitted**.

G.J.P. Van Westen, R.F. Swier, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Comparative Study and Benchmarking of 13 Amino Acids Descriptors and Applications to Proteochemometric Modeling.* **Submitted**.

G.J.P. Van Westen, O.O. van den Hoven, R. van der Pijl, T. Mulder-Krieger, H. de Vries, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data.* J. Med. Chem. *2012*. **55** (16): 7010-7020.

J.R. Lane, C. Klein Herenbrink, G.J.P. Van Westen, J.A. Spoorendonk, C. Hoffmann, and A.P. IJzerman; *A Novel Nonribose Agonist, LUF5834, Engages Residues That Are Distinct from Those of Adenosine-Like Ligands to Activate the Adenosine A2a Receptor.* Mol. Pharmacol.; *2012*. **81** (3): 475-487.

M.C. Peeters, Q. Li, G.J.P. Van Westen, and A.P. IJzerman; *Three "hotspots" important for adenosine A2B receptor activation: a mutational analysis of transmembrane domains 4 and 5 and the second extracellular loop.* Purinergic Signalling; *2012* **8** (1): 23-38.

G.J.P. Van Westen, J.K. Wegner, P. Geluykens, L. Kwanten, I. Vereycken, A. Peeters, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

G.J.P. Van Westen, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.

M.C. Peeters, G.J.P. Van Westen, Q. Li, and A.P. IJzerman; *Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation.* Trends Pharmacol. Sci.; *2011*. **32** (1): 35-42.

M.C. Peeters, G.J.P. Van Westen, D. Guo, L.E. Wisse, C.E. Müller, M.W. Beukers, and A.P. IJzerman; *GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor.* The FASEB Journal; *2011*. **25** (2): 632-643.

E. Van der Horst, J.E. Peironcely, G.J.P. Van Westen, O.O. Van den Hoven, W.R.J.D. Galloway, D.R. Spring, J.K. Wegner, H.W.T. Van Vlijmen, A.P. IJzerman, J.P. Overington, and A. Bender; *Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1964-1977.

G.J.P. van Westen, J.K. Wegner, A. Bender, A.P. IJzerman, and H.W.T. van Vlijmen; *Mining protein dynamics from sets of crystal structures using "consensus structures".* Protein Sci.; *2010*. **19** (4): 742-752.

M.R. Doddareddy, G.J.P. van Westen, E. van der Horst, J.E. Peironcely, F. Corthals, A.P. IJzerman, M. Emmerich, J.L. Jenkins, and A. Bender; *Chemogenomics: Looking at biology through the lens of chemistry.* Statistical Analysis and Data Mining; *2009*. **2** (3): 149-160.

Afterword

When pursuing a PhD for four years it is near impossible to describe everything and everyone that made an impact on you on a mere one or two pages. However, there is a large group of people who should be mentioned here as their help, support and co-operation was invaluable for the success of my PhD project.

First and foremost I would like to thank my two promoters Ad and Herman. Without both your vision and scientific input this would all have been a short exercise. Herman, it is incredible what effects a simple email, sent more than six years ago can have. Without the internship at Tibotec I am unsure if I would have continued to pursue a career in science. I think we can safely say that the science described in this thesis is a direct result from the way you supervised me during that time (remember our bi-weekly meetings). Ad I would like to thank you for the invaluable role you played in turning me from a cocky student into a proper scientist. Your patience, wisdom and in particular your skill at communicating one's results to others were something that will benefit me for the rest of my life. While we had clashes at times, I remember with pleasure the discussions we had and plans we made in both your office and mine. Working on a PhD project with two promoters like yourselves has been a very pleasant experience. This brings me to my co-promoter Andreas, whom has been another very significant influence on me these last years. It has been five years since we first met at the Gorlaeus restaurant, I remember our meeting as very energetic and focusing on opportunities and ideas. As a daily supervisor you showed the same qualities. Our day to day contact was always productive and our social outings were pleasant. Where I learned how to do science working with Ad and Herman, you taught me how to be a scientist.

Furthermore, I would also like to thank my MSc students, Olaf, Bart, Remco, Alwin and Marysa, and BSc student Tanja, for showing me the value of teaching and the crash course 'management in crisis situations' on a weekly basis. Your contributions have been invaluable and it is fun to see the direction each of you chose to go in after your biopharmaceutical sciences master. Also I should like to thank all the people from Medicinal Chemistry Leiden. Working at this department always felt very comfortable, but it was only at (inter)national conferences that I fully realized how fortunate I was to have worked in a group such as MedChem. The direct (and regular!) contact between experts with a diverse background (informatics, chemistry and bioassays) has been enriching. In particular I would like to thank Hans, Maris, Jaco and Laura for their expertise and help and Thea, Henk and Rianne for their massive contribution in performing all experimental work. Without you guys this thesis would also not have been completed.

I also would like to thank the members of my fraternity and in particular my co-founders Martin, Eric, Rob, Pouce, Caspar, Michel, Maarten, Gunn, GJ, Thijs en Jan (sorry Pouce I just cannot get myself to put Tom here). You guys showed me that indeed you can do anything you want; it's just a simple matter of perseverance. Furthermore, of all social circles I know you are one of the few where anybody can really be himself. In particular I would like to mention Michel. I have spent a total of 34 weeks receiving intravenous antibiotics these last years and you have joined me on nearly all of these trips to Amsterdam. Thank you for that, while you tend to systematically depreciate these actions (and any action that shows caring) it meant a lot to me. Thanks also for my friends with whom I enjoyed a 'cartoon avond' every week, Michael, Michel, Bastiaan and Dirk. These evenings were a very pleasant distraction from the life of a PhD student and a good place to brag and let off steam. Perhaps we can continue these events upon my return to the Netherlands. Special thanks go to Juriaan Bakx for almost 8 years of friendship and for showing me that you can actually do more valuable things than just playing videogames if you have a knack for computers. Our interests are overlapping to a large degree (ASOT) but also you were 'my guy on the inside' in the evil world of semi-corporate ICT at Leiden University. Another group of people that should be mentioned here are the 'Hufters'. The semi-regular gathering we enjoyed was always a good chance to speak to others 'in the trenches of PhD life'. I do hope we can keep the yearly outing alive as it does not get old (rather we do though).

Finally, I would like to thank both my parents for always supporting me and for, even though they disagreed on a course of action or had previous experience, allowing me to make my own mistakes to learn from. I would like to thank my mother for unconditional support but also for giving me her true opinion in every case. It is only now that I have just become a parent myself that I can truly value the way you have always been there. I can now fully understand why you asked the questions you did to my 3rd grade teacher. I would also like to thank my father, the role of a father can be difficult for some, but having had a very good example I must say this task befalls me easier than I feared beforehand. Also it was you who taught me the most valuable lesson of all when presenting something 'the audience doesn't know what you originally planned to say and hence will not judge you for that what you forgot to mention, just on what you actually presented'

Finally, I should like to mention those people who are most important and dear to me Afke, Max and Iris. Thank you for standing by me during everything that has happened over the last years. These last years have by no means been easy, not the PhD but my health was the most difficult hurdle. Thank you for understanding, supporting but mostly for the happy times we have together.

Curriculum Vitae

Gerard van Westen was born on March 28th 1983 in Leiden. He went to high school at the Stedelijk Gymnasium Leiden and graduated in 2001. In that same year he started his education at the Leiden/ Amsterdam Center for Drug Research (LACDR) in the undergraduate biopharmaceutical sciences. His first master internship was at the department of biopharmaceutics of the LACDR under supervision of Dr. Lutters and Prof. Dr. Biessen. Here he characterized the inhibition of P-Selectine using a peptide ligand. His second intership was at Tibotec BVBA (now Janssen pharmaceutica) in Mechelen (BE). Under supervision of Prof. Dr. Van Vlijmen and Dr. Wegner he started on a cheminfomatics and computational drug design project. During this internship Prof. IJzerman was his tutor at the LACDR.

After obtaining his master in Septermber 2007 he started as a PhD student in November of that year at the department of Medicinal Chemistry at the LACDR. In the period between November 2007 and March 2012 he performed the work described in this thesis. His PhD project was a cooperation between the LACDR and Tibotec, was funded by Tibotec, and his supervisors were Prof. Dr. IJzerman, Prof. Dr. Van Vlijmen and Dr. Bender.

During his PhD research he presented his work on multiple (inter)national conferences. He was invited as a speaker multiple times among others to the Dutch FIGON days (Lunteren, 2011) and to the Molsoft usergroup meeting (San Diego, 2012). At these conferences he was awarded multiple times for presentations and poster presentations.

He currently works as a postdoc at the European Bioinformatics Institute (part of the EMBL) in the ChEMBL group, headed by Dr. Overington, in Cambridge (UK). Gerard is married and has 2 children.

Curriculum Vitae

Gerard van Westen werd geboren op 28 maart 1983 te Leiden en groeide daar ook grotendeels op. Zijn middelbare schoolopleiding volgde hij aan het Stedelijk Gymnasium te Leiden, alwaar hij in 2001 zijn eindexamen afrondde. Vervolgens begon hij in datzelfde jaar aan de WO opleiding biofarmaceutische wetenschappen aan het Leiden/Amsterdam Center for Drug Research (LACDR). Zijn propedeuse behaalde hij in 2003 en zijn eindonderzoek startte hij in 2005 met een stage aan de afdeling Biofarmacie. Hier legde hij zich onder begeleiding van Dr. Lutters en Prof. Dr. Biessen toe op het karakteriseren van de inhibitie van P-Selectine door middel van een peptide ligand. In 2006 begon hij zijn tweede stage, welke plaats vond onder begeleiding van Prof. Dr. Van Vlijmen en Dr. Wegner bij het toenmalige Tibotec BVBA in Mechelen (BE, huidig Janssen Pharmaceutica) waarbij Prof. IJzerman zijn tutor aan het LACDR was. Deze stage richtte zich meer op cheminfomatics en computational drug design.

Na het behalen van zijn doctoraal examen in september 2007 begon hij in november van dat jaar als promovendus aan de afdeling Medicinal Chemistry van het LACDR. Tot maart 2012 verrrichtte hij het onderzoek wat in dit proefschrift beschreven staat onder begeleiding van Prof. Dr. IJzerman, Prof. Dr. Van Vlijmen en Dr. Bender. Zijn promotieproject was een samenwerking tussen het LACDR en Tibotec en werd gefinancierd door Tibotec.

Gedurende zijn promotie heeft hij op meerdere (inter)nationale en congressen zijn werk gepresenteerd. Hij is meermaals gevraagd als spreker onder andere op de FIGON dagen (2011) en op de usergroup meeting van Molsoft in San Diego (2012). Daarbij heeft hij meerdere prijzen gewonnen voor zowel presentaties en poster presentaties.

Vanaf mei 2012 werkt hij als postdoc aan het European Bioinformatics Institue (onderdeel van het EMBL) in de ChEBML groep, geleid door Prof. Dr. Overington, in Cambridge (VK). Gerard is getrouwd en heeft twee kinderen.

Appendix

Abbreviations

1D	–	One dimensional
2D	–	Two dimensional
3D	–	Three dimensional
AA	–	Amino Acid
ACE	–	Angiotensin-Converting Enzyme
AIDS	–	Acquired Immuno Deficiency Syndrome
AVG	–	Antivirogram
CCO	–	Clinical cut-off
CCP	–	Correctly Classified Percentage
CHO	–	Chinese Hamster Ovary
CPU	–	Central Processing Unit
CV	–	Cross Validation
DLV	–	Delavirdine
DT	–	Decision Tree
EL	–	Extracellular Loop
FASGAI	–	Factor Analysis Scales of Generalized Amino acid Information
FC	–	Fold Change
FN	–	False Negative
FP	–	False Positive
GP	–	Gaussian Processes
GPCR	–	G Protein-Coupled Receptor
GRIND	–	Grid Independent Descriptors
HAART	–	Highly Active Anti-Retroviral Therapy
HIV	–	Human Immunodeficiency Virus
LE	–	Ligand Efficiency
kNN	–	k-Nearest Neighbor
LOO	–	Leave-One Out
LOSO	–	Leave-One-Sequence Out
MCC	–	Matthews Correlation Coefficient
MHC	–	Major Histocompatibility Complex
MIPS	–	Million Instructions Per Second
NPV	–	Negative predictive value

NB	–	Naïve Bayesian
NN	–	Neural Net
NNRTI	–	Non-Nucleoside Reverse Transcriptase Inhibitor
NRTI	–	Nucleoside Reverse Transcriptase Inhibitor
NtRTI	–	Nucleotide Reverse Transcriptase Inhibitor
PCA	–	Principal Component Analysis
PCM	–	Proteochemometric
PC	–	Principal Component
PDB	–	Protein Data Bank
PI	–	Protease Inhibitor
PLFP	–	Protein-Ligand Fingerprint
PLS	–	Partial Least Squares
PPV	–	Positive predictive value
ProtFP	–	Protein Fingerprint
QSAR	–	Quantitative Structure-Activity Relationship
QSAM	–	Quantitative Sequence-Activity Modeling
RF	–	Random Forest
RMSE	–	Root Mean Squared Error
ROC	–	Receiver / Operator Characteristic
RS	–	Rough Set
RT	–	Reverse Transcriptase
Sens	–	Sensitivity
Spec	–	Specificity
SEM	–	Standard Error of the Mean
SVM	–	Support Vector Machines
TM	–	Trans Membrane
TN	–	True Negative
TP	–	True Positive
TEA	–	Two Entropy Analysis
VHSE	–	Vectors of Hydrophobic, Steric, and Electronic properties
VIP	–	Variable importance projection
VSS	–	Variable subset selection
Wt	–	Wild type

Glossary

Classification	–	Subtype of machine learning that predicts membership of any class present in the training set as output variable.
Compound	–	Chemical substance consisting of two or more different elements that can be separated into simpler substances by chemical reactions.
Data mining	–	The process of pattern discovery in large unsorted data sets.
Descriptor	–	Machine learning interpretable way of describing a compound or target.
External validation	–	Validation procedure for a statistical model based on pre-partitioning the data set with observations into two subdivisions ('Training set' and 'Test set'). Subsequently a statistical model is trained on the training set and validated on the test set.
False negative	–	A compound tested active but inactive according to a model.
False positive	–	A compound tested inactive but active inactive according to a model.
Floating point number	–	Computerized form of scientific notation consisting of the product of a mantissa and a power of 10 with the exponent expressed as an integer. However, floating point numbers can also use base 2, base 8, base 10 and base 16, where base 2 is the most common.
Internal validation/ Cross validation	–	Validation procedure for a statistical model based on partitioning of the training set into n subdivisions. Subsequently a statistical model is trained on n-1 subdivisions and validated on the remaining subdivision. This process is repeated n times.
Ligand	–	Compound that has been shown to bind to a protein of interest.
Negative predictive value	–	In classification, true negative compounds as fraction of the total of compounds inactive according to a model.

Positive predictive value	–	In classification, true positive compounds as fraction of the total of compounds active according to a model.
Regression	–	Subtype of machine learning that predicts a floating point number as output variable based on observed values found in the training set.
Sensitivity	–	In classification, true positive compounds as fraction of the total of compounds active according to experiments.
Specificity	–	In classification, true negative compounds as fraction of the total of compounds inactive according to experiments.
Small molecule	–	Organic compound with a molecular weight of under 500 Dalton
Target	–	Protein of interest in a medicinal chemistry project.
Training set	–	Collection of data points defined as observed examples to capture the distinction between desired compounds (e.g. ligands for a protein) and undesired compounds.
True negative	–	Compound tested inactive and inactive according to a model.
True positive	–	Compound tested active and active according to a model.
Test set	–	Collection of data points used to validate a trained statistical model before production usage.

List of figures

List of tables